PROJECT PROPOSAL

IE7500 | NEU

NLP Sentiment Analysis

Apply engineering principles to the NLP problem of Sentiment Analysis

January 21, 2025

Group E

Arundhati Ubhad, Jensen Ho, Harjot Virk

Sentiment Analysis of E-Commerce Product / Hotel Reviews

This project aims to build a machine learning pipeline that automatically classifies product reviews into positive, negative, or neutral sentiments. By preprocessing the raw text, extracting relevant features, and training a classification model, we will demonstrate how computational methods can be leveraged to interpret and label textual data at scale. Ultimately, this system could help e-commerce platforms and businesses understand customer feedback in real time, informing improvements to products and services.

Problem Statement

The core challenge lies in accurately determining whether a customer's review conveys a positive, negative, or neutral sentiment. This is a fundamental Natural Language Processing (NLP) problem, as it involves interpreting human language patterns—tokenizing text, handling various linguistic nuances, and extracting meaningful features from words and phrases. The solution not only aids in automating feedback analysis but also illustrates important NLP concepts such as text representation and classification. We think that classification can be used to understand sentiment analysis results with respect to keywords.

Dataset Selection

Source:

We will use this publicly available dataset from Kaggle with 568454 records:
https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/data
Reviews from Oct 1999 - Oct 2012
568,454 reviews
256,059 users
74,258 products
260 users with > 50 reviews

Structure:

Each entry includes Id, ProductId, UserId, ProfileName, HelpfulnessNumber, HelpfulnessDenominator, Score, Time, Summary, Text

Volume:

The datasets contain over 500,000 reviews, providing a rich corpus for sampling, training, and evaluating the sentiment classifier. We will be creating random 50000 sample sets for training. Sample size of 50000 ensures the model can learn from diverse language styles and product contexts.

Tasks for the team:

Sampling dataset
Research paper summarization and parameter selection
Data preprocessing for correction, and cleaning.

Expected Outcomes

Model:
A trained sentiment classification model capable of categorizing reviews into positive, negative, or neutral sentiment with a reasonable degree of accuracy. Three types of models will be investigated for suitability including traditional machine learning, deep learning, and transformer.

Evaluation:
The performance will be assessed using metrics such as accuracy, precision, recall, and a confusion matrix to pinpoint areas for further improvement. Check for true positives, true negatives, false positives, and false negatives.

Insights:
The project will highlight best practices in NLP—covering data cleaning, feature engineering, and model selection—while offering actionable insights into customers' perceptions of products.