NIT Rourkela

Lab-3, CS6379

NLP Lab, Spring'25

Topics: Sentiment/Classification analysis

Problem Statement:

1. You are given a dataset containing a collection of emails labeled "Spam" or "Not Spam." Your task is to build a **Naive Bayes classifier** that can predict whether a new email is spam or not based on its content.
2. You have a dataset containing articles from different domains, such as **Sports, Politics, and Technology**. Your task is to train a **Naive Bayes classifier** that can automatically categorize a new document into one of these domains based on its content.
3. You are given a dataset containing movie reviews labeled as **Positive** or **Negative**. Your task is to build a **Naive Bayes classifier** to predict the sentiment of a new review.
4. You have a dataset of **customer product reviews** labeled as **Positive, Neutral, or Negative**. Your goal is to develop a **Naive Bayes classifier** that can predict the sentiment of a given product review.

NOTE:

1. Email Spam Detection:
   - A dataset containing emails labeled as Spam or Not Spam.
   - Each email consists of raw text data.
2. Document Classification:
   - A collection of documents labeled as Sports, Politics, Technology, etc.
   - Each document contains textual content related to its category.
3. Movie Review Sentiment Analysis:
   - A dataset of movie reviews labeled as Positive or Negative.
   - Each review is a piece of raw text expressing sentiment.
4. Product Review Sentiment Analysis:
   - A dataset containing customer product reviews labeled as Positive, Neutral, or Negative.
   - Reviews are textual descriptions of product experiences.

5. Use **TF-IDF** for numerical representation of text.

6. Split the dataset into **training and test sets**. Use **accuracy, precision, recall, and F1-score** to measure performance. Train the model and apply it to unseen data.