**Machine learning and Data Science Laboratory**

ASSIGNMENT: 6

Title: Sentiment Analysis for Document Polarity Classification

Aim:

The aim of this project is to develop a document polarity classification system that can automatically recognize whether a given document has a positive or negative sentiment. This system will be based on the analysis of polarity words and will be a valuable tool for businesses and individuals seeking to quickly assess the sentiment of textual data.

Objectives:

1. Collect and preprocess a dataset of documents with labeled positive and negative sentiments.

2. Identify and extract polarity words from the dataset, which are words indicative of positive or negative sentiment.

3. Implement a classification method to predict document sentiment using the extracted polarity words.

4. Evaluate the model's performance through metrics such as accuracy, precision, recall, and F1-score.

5. Develop a user-friendly interface to input documents and receive sentiment predictions.

Theory:

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) task that involves determining the sentiment or emotional tone expressed in a piece of text. In this project, we focus on binary sentiment analysis, where documents are classified as either positive or negative.

The classification method will be based on machine learning techniques, specifically a supervised approach. The following steps will be taken:

1. Data Collection and Preprocessing:

- Gather a labeled dataset of documents with positive and negative sentiments.

- Preprocess the data, including text cleaning, tokenization, and removing stopwords.

2. Polarity Words Extraction:

- Identify and extract polarity words from the dataset. These are words that strongly indicate positive or negative sentiment, such as "happy" for positive and "unhappy" for negative sentiment.

3. Feature Extraction:

- Convert the documents into numerical feature vectors using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings.

4. Classification Model:

- Choose a suitable machine learning algorithm, such as Naive Bayes, Support Vector Machines (SVM), or deep learning models like LSTM or BERT.

- Train the model using the labeled dataset and the extracted features.

5. Model Evaluation:

- Evaluate the model's performance using appropriate metrics, including accuracy, precision, recall, and F1-score.

- Fine-tune the model parameters for optimal performance.

6. User Interface:

- Develop a user-friendly interface where users can input documents and receive sentiment predictions.

Conclusion:

This project aims to create an efficient document polarity classification system using machine learning and sentiment analysis techniques. By extracting and analyzing polarity words, the model will provide valuable insights into the sentiment of textual data. The system will be beneficial for various applications, including market research, customer feedback analysis, and social media monitoring, enabling automated sentiment assessment for decision-making processes. The success of this project will depend on the accuracy and reliability of the classification model, as well as the user-friendliness of the interface.