**Natural Language Processing (NLP) Sentiment Analysis on Employee Feedback**

Sentiment analysis put simple is a technique used to determine whether text data is positive, neutral, or negative.

**Steps for Sentiment Analysis**:

**Step 1: Data Collection**

Collect Feedback Data: Gather employee feedback from surveys, reviews, or any relevant source. This data should be in a structured format, such as a CSV or Excel file, where each feedback entry corresponds to an employee or job title.

Ensure Diversity: Collect feedback from different levels of experience, job titles, and departments to get a comprehensive view.

**Step 2: Text Preprocessing**

Text preprocessing is essential to clean and prepare your data for analysis. Common preprocessing steps include:

Lowercasing: Convert all text to lowercase to maintain consistency.

Removing Punctuation: Eliminate punctuation marks to focus on words.

Tokenization: Split the text into individual words or tokens. This can be done using simple text manipulation techniques.

Removing Stop Words: Remove common words that do not carry significant meaning (e.g., "and," "the," "is"). You can use pre-defined lists of stop words.

Stemming/Lemmatization: Reduce words to their root form (e.g., "running" to "run"). This step helps in standardizing words.

**Step 3: Feature Extraction**

Convert the cleaned text data into numerical representations to prepare for modeling. You can use:

Bag of Words (BoW):

Count the frequency of each word in the feedback.

Create a matrix where each row represents an employee's feedback, and each column represents a unique word from the entire corpus.

TF-IDF (Term Frequency-Inverse Document Frequency):

Weigh words based on their frequency in a document relative to their frequency across all documents.

This method highlights important words in feedback while downplaying commonly used words.

**Step 4: Modelling**

Choose a Sentiment Analysis Model: Common algorithms for sentiment analysis include:

Naive Bayes: A simple and effective algorithm for text classification.

Logistic Regression: Suitable for binary classification tasks.

Support Vector Machines (SVM): Effective for high-dimensional data.

Train the Model: Use your feature-extracted data to train the model. Split your data into a training set (e.g., 80%) and a test set (e.g., 20%) to validate model performance.

**Step 5: Model Evaluation**

Evaluate Performance: After training the model, assess its performance using various metrics:

Accuracy: The percentage of correct predictions out of total predictions.

Precision: The proportion of true positive predictions relative to the total predicted positives.

Recall: The proportion of true positives relative to the total actual positives.

F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.

Use Confusion Matrix: A confusion matrix can help visualize how well your model performed in classifying feedback into positive, negative, and neutral categories.