Experiment No.1
Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications
Date of Performance:
Date of Submission:

Aim: Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications.

Objective: Understand the different applications of NLP and their techniques by reading and critiquing IEEE/ACM/Springer papers.

Theory:

1. **Machine Translation**

Machine translation is a process of converting the text from one language to the other automatically without or minimal human intervention.

2. **Text Summarization**

Condensing a lengthy text into a manageable length while maintaining the essential informational components and the meaning of the content is known as summarization. Since manually summarising material requires a lot of time and is generally difficult, automating the process is becoming more and more popular, which is a major driving force behind academic research.

Text summarization has significant uses in a variety of NLP-related activities, including text classification, question answering, summarising legal texts, summarising news, and creating headlines. Additionally, these systems can incorporate the creation of summaries as a middle step, which aids in shortening the text.

The quantity of text data from many sources has multiplied in the big data era. This substantial body of writing is a priceless repository of data and expertise that must be skillfully condensed in order to be of any use. A thorough investigation of NLP for automatic text summarization has been necessitated by the increase in the availability of documents. Automatic text summarising is the process of creating a succinct, fluid summary without the assistance of a human while maintaining the original text's meaning.

3. Sentiment Analysis

Sentiment analysis, often known as opinion mining, is a technique used in natural language processing (NLP) to determine the emotional undertone of a document. This is a common method used by organisations to identify and group ideas regarding a certain good, service, or concept. Text is mined for sentiment and subjective information using data mining, machine learning, and artificial intelligence (AI).

Opinion mining can extract the subject, opinion holder, and polarity (or the degree of positivity and negative) from text in addition to identifying sentiment. Additionally, other scopes, including document, paragraph, sentence, and sub-sentence levels, can be used for sentiment analysis.

Businesses must comprehend people's emotions since consumers can now communicate their views and feelings more freely than ever before. Brands are able to listen carefully to their customers and customise their products and services to match their demands by automatically evaluating customer input, from survey replies to social media chats.

4. Information Retrieval

A software programme that deals with the organisation, storage, retrieval, and evaluation of information from document repositories, particularly textual information, is known as information retrieval (IR). The system helps users locate the data they need, but it does not clearly return the questions' answers. It provides information about the presence and placement of papers that may contain the necessary data. Relevant documents are those that meet the needs of the user. Only relevant documents will be pulled up by the ideal IR system.

5. Question Answering System (QAS)

Building systems that automatically respond to questions presented by humans in natural language is the focus of the computer science topic of question answering (QA), which falls under the umbrella of information retrieval and natural language processing (NLP).

CSE-DS
Dikshant Buwa - 4
Mayank Jadhav-20
Yash Sankhe-53
Arpit Sutariya-59

XGBoost-based Spam Classification System: Improving Email Security

Abstract:

Spam messages, or unsolicited and unwanted texts, pose significant challenges in maintaining the integrity of digital communication channels. To combat this issue, automated spam message detection systems have become crucial for protecting users from malicious content and preserving the efficiency of communication platforms. In this study, we propose a robust text classification approach using XGBoost, a powerful gradient boosting algorithm, to identify and classify spam messages effectively. Our method leverages the XGBoost algorithm's ability to handle high-dimensional data and nonlinear relationships, making it suitable for text classification tasks. As a feature extraction technique, we employ CountVectorizer to convert raw text messages into numerical vectors, representing the frequency of words in the text. This step ensures the compatibility of the data with the XGBoost algorithm.

Methodology:

1. Data Collection: Obtain a dataset containing a collection of text messages labeled as spam or non-spam (ham). Dataset collected from Kaggle.

2. Data Preprocessing: Remove any irrelevant information, such as special characters, symbols, or HTML tags, from the messages.Handle any missing or null values in the dataset.

3. Data Splitting : Split the dataset into training and testing sets to evaluate the model's performance accurately. Typically, use a 70-30 or 80-20 split for training and testing, respectively.

4. XGBoost Model Training: Initialize the XGBoost classifier with suitable hyperparameters. Train the XGBoost model on the training data, using the feature vectors and corresponding labels (spam or non-spam). During training, the model will build an

ensemble of decision trees to learn the relationship between the feature vectors and the target labels.

5. Model Evaluation: Use the trained XGBoost model to predict the labels of the test data (messages).Evaluate the model's performance using various metrics, such as accuracy.

Process (Boosting algorithm):

1. Pandas: A Python library used for data manipulation and analysis. It provides data structures and functions needed to work with structured data, like CSV files, in a way that is both efficient and easy to use.

2. Seaborn: A data visualization library built on top of Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

3. Matplotlib: A data visualization library in Python that provides a flexible way to create various types of plots and charts.

4. NumPy: A fundamental package for numerical computations in Python. It provides support for large, multi-dimensional arrays and matrices.

5. Train-Test Split: The process of dividing the dataset into training and testing subsets. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance.

6. Accuracy Score: A metric used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances out of the total instances.

XGBoost Algorithm:

XGBoost (Extreme Gradient Boosting) is a popular and powerful gradient boosting algorithm used for both regression and classification tasks. It is an ensemble learning method that combines multiple weak learners (decision trees) to create a strong learner.