**ICP2 WIKI REPORT**

**Description:**

In the given ICP We have to perform the sentiment analysis task on the given data using one of the scikit learn classifiers for text.

**Objective:**

The main objective of ICP is to perform sentiment analysis and data cleaning and preprocessing, removing unnecessary columns or data, removing Twitter Handles(@user), removing punctuation, numbers, special characters, removing stop words, Tokenization, and Stemming, TFIDF vectors, POS tagging, checking for missing values, train/test split of data), data visualization and analysis for critical steps (WordCloud, Bar plots, etc.)

**Design and Implementation:**

1. To perform this task, we need to import all the below-required libraries

Text

Description automatically generated

1. Importing the Data from CSV

Graphical user interface, text, application, email

Description automatically generated

1. Generating Descriptive statistics include those that summarize the central tendency, dispersion, and shape of a dataset's distribution, excluding NaN values

Graphical user interface, text, application

Description automatically generated

1. Checking For Null Values present or not

A picture containing table

Description automatically generated

1. By using Regular expression and NumPy library (faster than python lists) we are removing Twitter Handles (@user) from the 'tweet' column and storing it into another column 'Tweets'

Graphical user interface, text, application

Description automatically generated

1. Removing punctuation, numbers, special characters by using regular Expression and also removing unwanted column tweets by using drop()

A picture containing table

Description automatically generated

1. With the help of nltk.word\_tokenize () method, we are able to extract the tokens for Tweets

Chart

Description automatically generated

1. By using POS tagger assigning parts of speech, grammatical information for each word of the sentence.

Text, letter

Description automatically generated

1. By using FreqDist() getting the frequency of words within a Tweets

Graphical user interface, text, application

Description automatically generated

1. By using Fdist.Plot() from the frequency distribution displaying the most frequent words in the Tweets.

Chart, line chart

Description automatically generated

1. By using stopwords.words("english") we are removing stopwords from Tweets.

Graphical user interface, text, application

Description automatically generated

1. By using Fdist.Plot() from the frequency distribution displaying the clean words after removing stopwords in the Tweets

Text

Description automatically generated

Chart, line chart

Description automatically generated

1. By using Wordcloud displaying a visual representation of clean words. Words are usually single words, and the importance of each is shown with font size or color

Graphical user interface, text, application

Description automatically generated Text

Description automatically generated

1. By using Stemming with Python nltk package reducing inflection in words to their root forms.

Text, letter

Description automatically generated

1. By using WordNetLemmatizer() we are getting all the lemmas for Tweets.

Graphical user interface, text, application, email

Description automatically generated

1. By using train\_test\_split function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data

Text

Description automatically generated

1. By using Wordcloud displaying visual representation for Positive\_words where label==1.

Graphical user interface, text

Description automatically generated with medium confidenceText

Description automatically generated

1. By using Wordcloud displaying visual representation for Negative\_words label==0

Graphical user interface, text, application, email

Description automatically generatedText

Description automatically generated

1. By using TfidfVectorizer get word frequency scores.

Table

Description automatically generated with medium confidence

1. By using sns.countplot() displaying bar graph for labels.

Chart, bar chart

Description automatically generated

Video Link: <https://drive.google.com/file/d/1Zcb4YrPI5HE2-w1Jv_nNrHtRrJtiJ5w1/view?usp=sharing>

**Conclusion:**

**Learning from the lesson:**

we have learned how to perform sentiment analysis, build word clouds, beautiful visualizations, and machine learning models using text data.

**Issues with the lesson:**

I have faced some issues while performing TFIDF vectors and conversions of data into different data types.