horizontal line



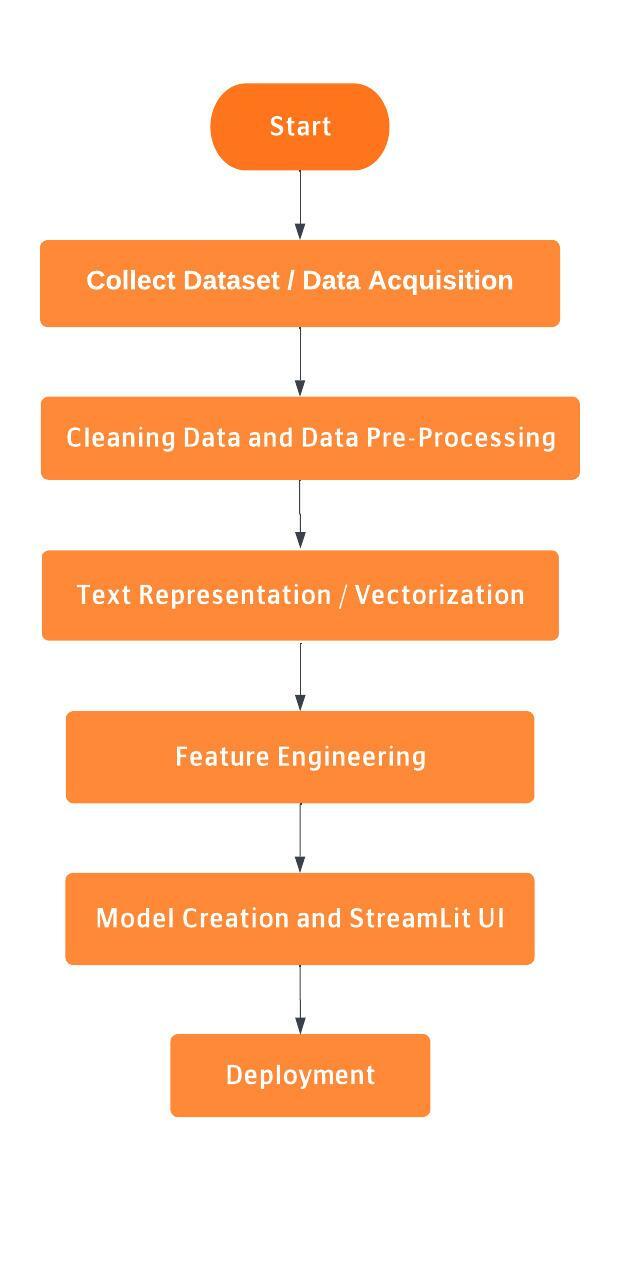
Democratizing NLP - A No Code Framework

**Nov 11, 2022**

# GOALS OF THE PROJECT

* Provide an interface where a User can upload the Dataset and perform various Natural Language based tasks on the Dataset
* The User can select the Pre-Processing Methodologies to be advocated, and can download the Pre-Processed Dataset for further computation
* The User can perform Machine Learning Algorithms on the Dataset, and can determine the accuracy of the established model
* A total of 5 ML Algorithms have been taken - KNN, Logistic Regression, SVM, Random Forest and Naive Bayes, which are all Classification Algorithms
* Any model, which the User decides, can be downloaded as a pickle file, and can be used for further computation
* A feature named “NER as a Text Box” has also been included with the Project

# PROJECT PIPELINE



# STEPS TO RUN THE MODEL

* Download the Zip File and extract the contents
* Create a virtual environment preferably anaconda
* Navigate to the directory from the terminal
* Run the Command:

conda env create -f env\_name

* Activate the Environment

conda activate env\_name

* Download the Requirements

pip install -r requirements.txt

* Run the Streamlit UI

streamlit run application.py

# NATURAL LANGUAGE PROCESSING DONE IN MODEL

# STANDARD PRE-PROCESSING

## Text Cleaning

The Dataset may contain unwanted data. Therefore, all the characters will be converted to lowercase, and the punctuation characters will be removed.

* **Stop Words Removal**

Stop Words are the words which are most commonly used in any language, and they do not provide any contribution to the meaning of the sentence. Therefore, these stop words are removed from the dataset for giving more focus to the important words.

* **Removal of Null Values**

If the Dataset contains any null values within the data, the entire row will be deleted so as to not create errors while processing the dataset.

* **Removal of Duplicate Values**

If the dataset contains any row which is exactly same as another row in the dataset, then only one of these rows is retained, and the other rows are deleted.

* **Tokenization**

Tokenization of the corpus can be done in 2 ways : Word Tokenizer and Sentence Tokenizer. According to the input given by the user, the corpus can be tokenized in either of the two ways. Named Entity Recognition (NER) can be used for obtaining a better result in Tokenization. NER can be used to extract important entities such as names of People, Location and Organizations.

* **Text Normalization**

Converting Tokens into Standard Form. We use both Stemming and Lemmatization. These operations reduce the volume of the text. The User can choose which methodology to use, or can select both.

* **Stemming**

The suffixes from the words are removed to give the stem word. The stem words are inflected with the affixes to produce the given word. Porter Stemmer, Snowball Stemmer and Lancaster Stemmer are few commonly used Stemming Algorithms. The problem with Stemming is that there are words which do not have a proper stem word after removal of the suffix, therefore, they are meaningless.

# OTHER PRE-PROCESSING STEPS (Given as Menu)

* Remove Digits
* Remove Alphabets
* Remove AlphaNum
* Only retain AlphaNum
* Remove Special Characters

The User has been provided with the option of selecting one or more Pre-Processing Standards in the Model.

**DATA ANALYSIS DONE IN NLP PROJECTS**

Python provides many packages for visually representing the data. Therefore, it really helps which doing Exploratory Data Analysis on the Dataset.

**Some of the Data Analysis methods are:**

* **Word Frequency Analysis**

Using barplots, we can represent the frequency of each feature using bars. Therefore, we can compare the frequencies of different features.

The average length of each row of each feature can also be represented using bar plots.

* **Stop Word Analysis**

The average occurrence of stop words can be calculated and represented using bar plots. If the stop words count is high, then the dataset can be reconsidered for applying ML algorithm.

* **Distribution of target variable**

Distribution of Positive Outcomes and Negative Outcomes are calculated and presented as bar graphs in order to check if there is equal distribution.

* **Python Word Cloud**

For Text Data, Word Cloud is a good way for representing Textual Data. For this, the data will be converted into a corpus. The size and color of each word that appears in the wordcloud indicate it’s frequency or importance.

Therefore, since there is no hard and fast rule for Exploratory Data Analysis, a number of operations can be performed on the Dataset.

**MACHINE LEARNING ALGORITHMS USED**

A total of 5 ML Algorithms namely - KNN, Logistic Regression, Naive Bayes, SVM and Random Forest have been used.

The accuracy of each algorithm can be represented by using the StreamLit application, and the best accurate model can be downloaded as a pickle file. The download request is placed as a button in the application.

**Project Mentor**