# *FAKE NEWS DETECTION*

# *USING NLP*

# *ARITIFICAL INTELLIGENCE PROJECT*

## *STATEMENT:*

***What Is Fake News Detection Using Machine Learning Project?***

***Detecting fake news using machine learning techniques would mean having an automatic detection system that looks at a piece of text (tweets, news articles, a WhatsApp message) and determine how likely it looks like a piece of false news. The system will be a machine learning model trained on a large enough dataset containing examples of real and false news from various sources and styles. Machine learning models performing binary classification can be trained on this training set. However, since machine learning models only look at numerical features, we must perform natural language processing on this text corpus (collection of text samples).***

***Natural language processing will perform data cleaning, stemming, lemmatization, and vectorization using one of the many available techniques and convert sentences into a vector of numbers that machine learning models can interpret. Once this is done, we can train models like Naive Bayes, Logistic Regression, and Random Forests and observe their results.***

***If we find that the performance of these machine learning techniques is lacking in the dataset, we can delve into deep learning and look at LSTM or Attention-based models to perform text classification.***

***But first, let us see why you should use machine learning for detecting false news and what drawbacks you should be aware of while doing***


## *DESIGN THINKING:*

## *1. DATA SOURCE:*

***TOP5 Fake News Detection Project Datasets:***

***Since the past few years, as the spread of online fake news has become more and more of a problem, the word appears that many publicly available datasets contain news articles, tweets, or posts along with an appropriate label for their legitimacy.***

***FakeNewsNet: Dataset of Political and Gossip Tweets:***

***This dataset uses the Twitter official API to fetch tweets from Twitter users, including metadata and social context. But to begin with, they provide a sizeable amount of clean data that can be readily used to test models.***

***Fake News Corpus:***

***This open-source dataset comprises millions of news articles scraped from a curated list of 1001 domains. The dataset includes over 9,400,000 articles from over 700 domains scraped from multiple domains, such as NYTimes and WebHose English News Articles.***

***FakeHealth:***

***FakeHealth is collected to address Fake Health News detection challenges from sources like news content, reviews, and social media. As proposed in the paper, "Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository," the data can be fetched from Twitter using the Twitter API and the IDs provided in the GitHub repository- EnyanDai/FakeHealth. Ultimately, the dataset contains data from 500k tweets, 29k replies, 14k retweets, and 27k user profiles with timelines and friend lists. The reviews cover explanations regarding ten health news evaluation criteria.***

***Constraint COVID-19 Fake News Dataset:***

***This dataset contains social media posts related to COVID-19 and vaccinations from various popular platforms. The posts are annotated as real and fake news, and the dataset can be better understood in its paper 'Fighting an Infodemic: COVID-19 Fake News Dataset' and the data, as well as the proposed baseline benchmark using attention models and heuristic-based post-processing, can be found in the GitHub repository- COVID Fake News***

***Social Media:***

***Fake news spreads quickly on social media platforms, leading to misinformation and confusion. Many leading social media platforms have implemented false news detection algorithms to combat this issue. For example, Twitter uses machine learning to detect and flag potentially misleading or false information in tweets.***

***News/Journalism***

***News organizations use machine learning algorithms to verify information and sources. For example, the BBC developed a tool called "Project Origin," which uses natural language processing (NLP) and machine learning techniques to identify the origin of images and videos and determine if they are authentic or have been manipulated.***

***Politics:***

***Fake news can significantly impact political campaigns and elections. Several political organizations have developed tools to detect false news and misinformation. For example, the German political party CDU developed a tool called "Augmented Intelligence for Quality Journalism" (AIQ), which uses machine learning to identify and fact-check news articles.***

***Finance:***

***Fake news can also significantly impact financial markets. Hedge funds and investment firms use machine learning algorithms to analyze news sources and detect false news that could impact the stock market. For example, Dataminr uses machine learning techniques to analyze social media feeds and news sources to provide real-time alerts on events that could impact financial markets.***

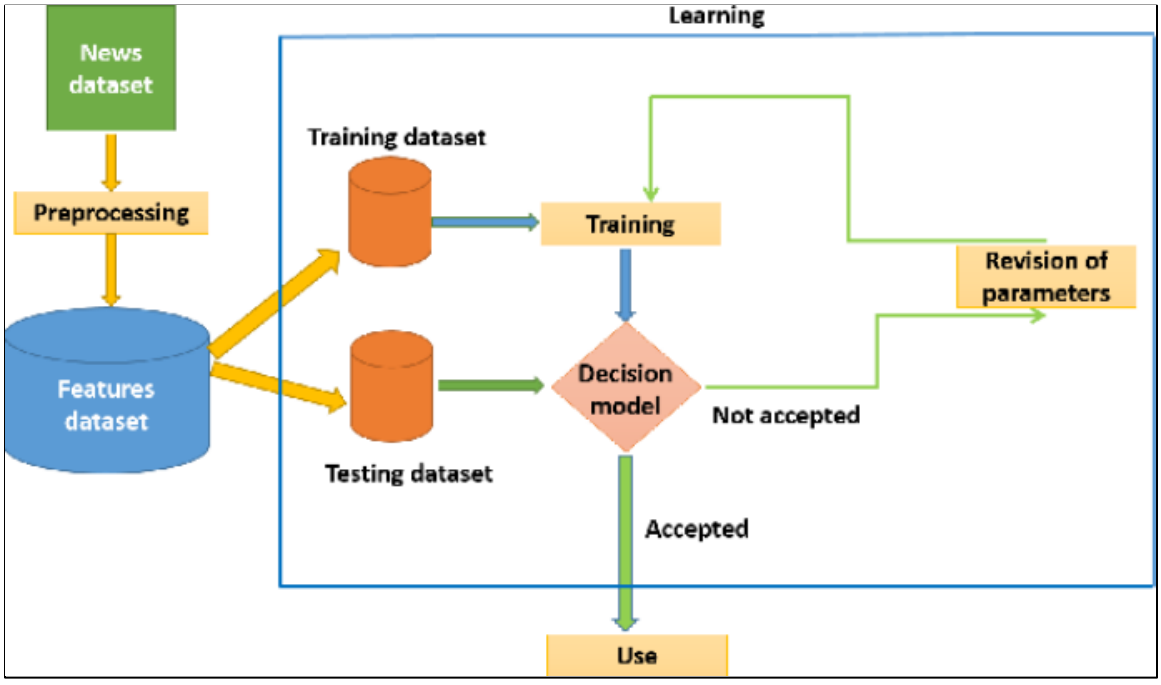
***Healthcare:***

***Fake news can also seriously affect the healthcare industry. Several organizations use machine learning to identify false healthcare and medical research news. For example, researchers at the University of Michigan developed an IRIS tool that uses natural language processing and machine learning to identify fake news related to healthcare and medical researce***

***2. DATA PREPROCESSING:***

***To preprocess your text simply means to bring your text into a form that is predictable and analyzable for your task.***

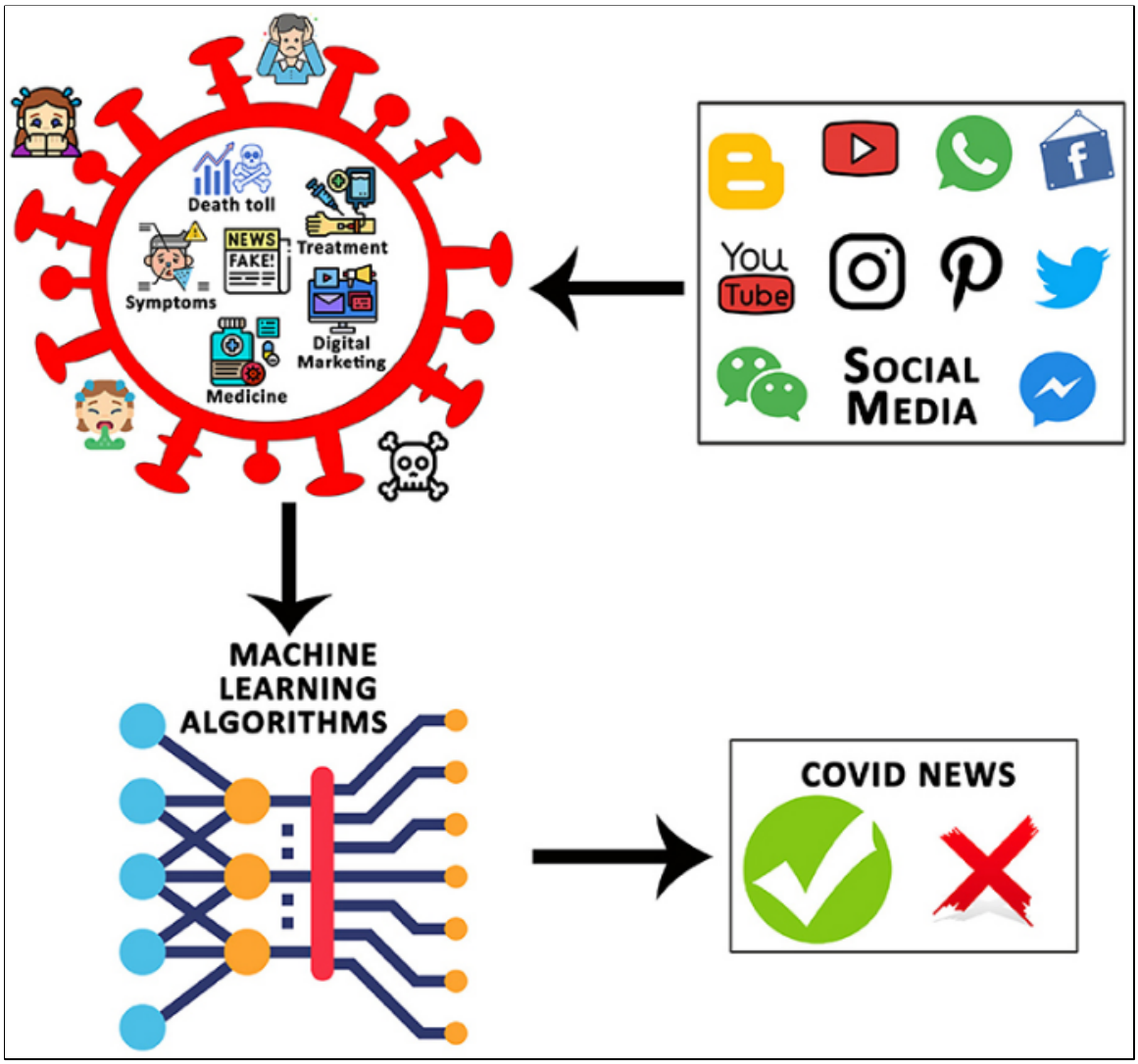
***The goal of pre-processing is to remove noise. By removing unnecessary features from our text, we can reduce complexity and increase predictability (i.e. our model is faster and better). Removing punctuation, special characters, and filler word***




***3.FEATURE EXTRACTION:***

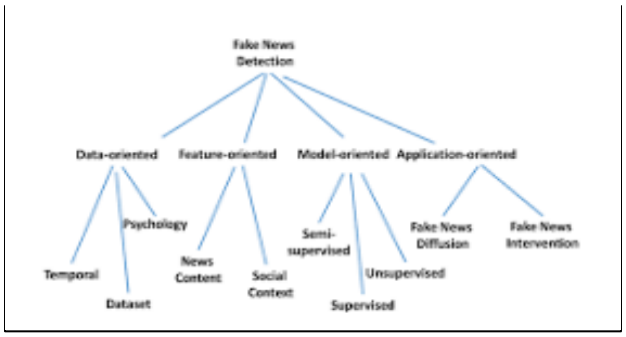
***The major contribution of this work is the extraction of important features from the COVID-19 fake news dataset. Feature extraction plays a very important role in text***

***processing as it reduces the dimension of feature space by considering only the important features (27–29). To extract the features, the named-entity recognition (NER) approach is used in our work. The NER is a popular approach for feature extraction that can classify unstructured text based on location, person names, quantities, etc.***




***4.MODEL SELECTION:***

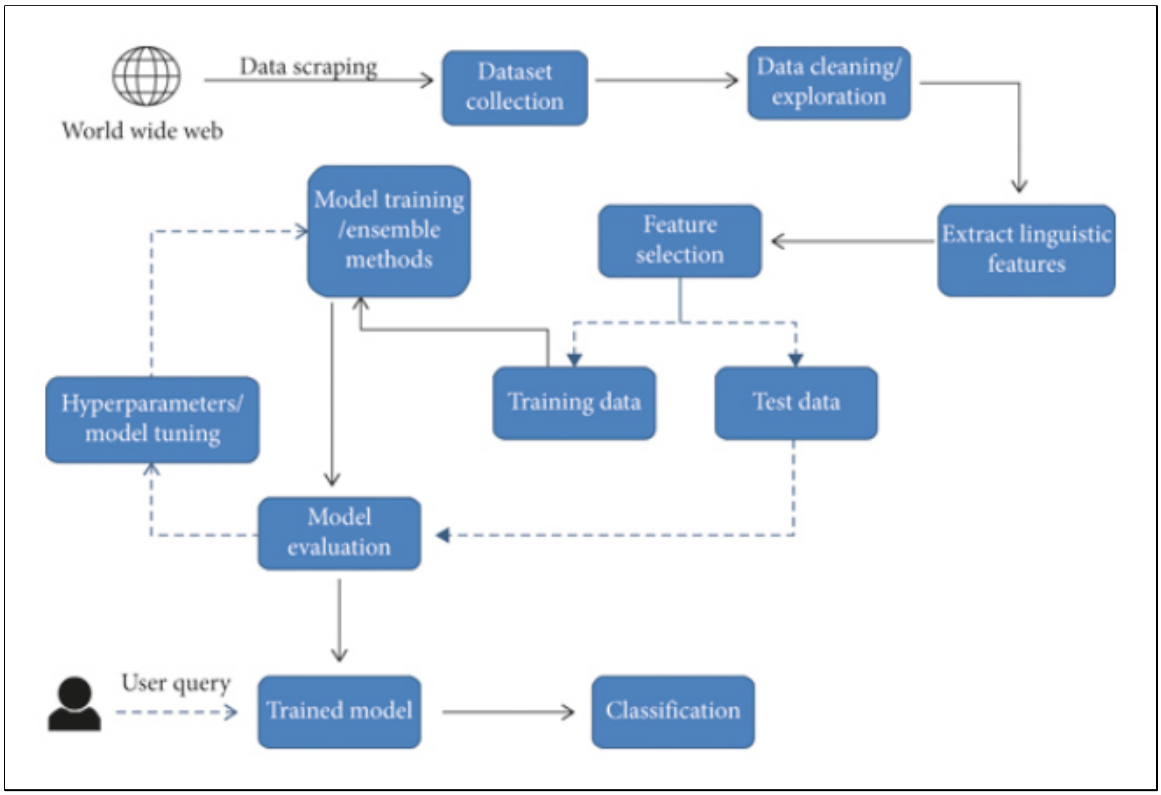
***A classification column needs to be added to identify whether a record is fake news or not. With the command fake['classification'] = 1 we create a new column called classification, and all its rows are set equal to 1. Our classifier will identify fake news and assign them the value of 1.24***




***5.MODEL TRAINING:***

***detecting fake news with Python, start by first preprocessing the input text, getting numerical features, and then training a machine learning model like SVM, LSTM, or an RNN to predict whether the news is reliable or not***

***To detect false news, you must first preprocess the input text data to convert it into a numerical representation. Then, you must use this representation as input to the MLP, which is trained to classify the news as fake or genuine based on the features extracted from the input text data.***



***RESULT:***

***Performance Metrics (Accuracy, Precision,Recall,F1-score)***

***Confusion Matrix***

***Visulization of Results***

***CONCLUSION:***

***Summary of key findings***

***Importance of Fake News Detection***