Fake news detection using Natural language processing

Date	29 September 2023
Team ID	Proj_212171_Team 1
Project name	Fake news detection using NLP
Maximum marks	

Design Thinking Process

The five stages of design thinking process are:

Empathize

research your users' needs.

Define:

state your users' needs and problems

Ideate:

challenge assumptions and create ideas.

Prototype:

start to create solutions.

Test:

try your solutions out.

Let’s dive into each stage of the design thinking process.

Stage 1: Empathize—Research Your Users' Needs

We consume news through several mediums throughout the day in our daily routine, but sometimes it becomes difficult to decide which one is fake and which one is authentic.

Every news that we consume is not real. If you listen to fake news it means you are collecting the wrong information from the world which can affect society because a person’s views or thoughts can change after consuming fake news which the user perceives to be true.

Stage 2: Define—State Your Users' Needs and Problems

In the Define stage, you will organize the information you have gathered during the Empathize stage. You'll analyze your observations to define the core problems you and your team have identified up to this point.Defining the problem **and problem statement must be done in a human-centered manner**.

Requirements:

Things you need to install

Python 3.9

This setup requires that your machine has python 3.9 installed on it. you can refer to this url https://www.python.org/downloads/ to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/.

You will also need to download and install the required packages after you install python

- Sklearn (scikit-learn)
- numpy
- Pandas
- matplotlib
- seaborn
- NLTK
- Joblib

To install the Packages

- pip3 install -U scikit-learn
- pip3 install numpy
- pip3 install Pandas
- pip3 install matplotlib
- pip3 install seaborn
- pip3 install nltk
- pip3 install flask
- pip3 install joblib

Dataset

All of the Dataset that used in this project are availabe in public Domain.Most of the Dataset are collected from Kaggle (https://www.kaggle.com/) different datsets contain different column and different information like [title,text,subject,news_url,author]

For model Build need only text and Label,The final dataset will contain only 2 column ['Article','Lable']

For text we will create a news column named 'Article' which is the Combination Header and text

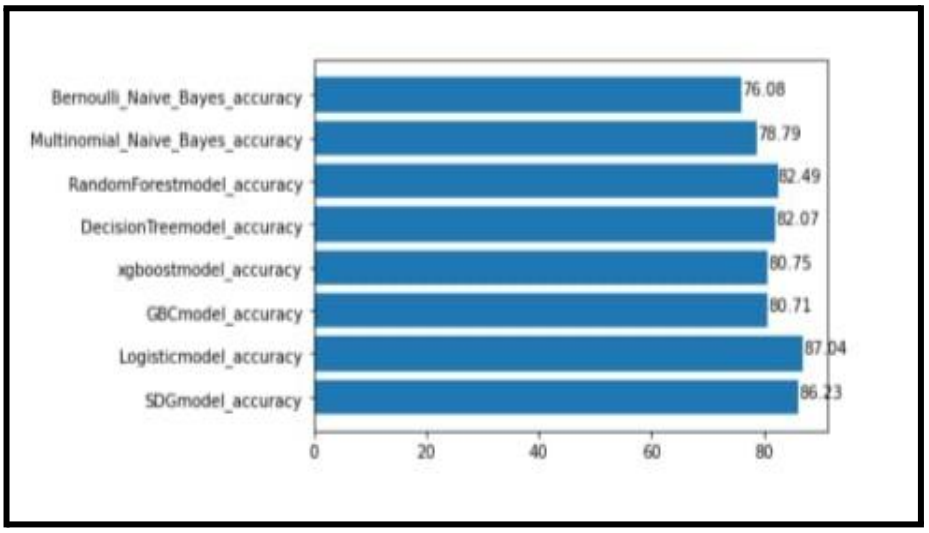
In the Lable column

- 1 represent true
- 0 represent fake

Stage 3: Ideate—Challenge Assumptions and Create Ideas

Ideate: the third phase of design thinking, where you identify innovative solutions to the problem statement you've created.

ML model Traning and Building for Detection of News whether fake or not
Here we have build all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. We have used Logistic Regression, Stochastic gradient descent,Random forest, GBC, xgboost, DecisionTree, Multinomial Naive Baye and Bernoulli Naive Baye classifiers . Each of the extracted features were used in all of the classifiers. Once fitting the model, we compared the accuracy score and checked the confusion matrix.



Stage 4: Prototype—Start to Create Solutions

Prototype: the fourth phase of design thinking, where you identify the best possible solution.

Methodology

1)CountVectorizer

2) TF-IDF

3)LSTM

20800 train and 5200 test news dataset used to classify the fake and real news using Count Vectorizer and TF-IDF. Seven ML algorithms are applied to find the best model for the dataset.

CountVectorizer:

1.Taking data from kaggle of 20800 data

2.Preprocessing : Remove RE, special character, remove stop words, make all lower case

3.USe bag of words method to make feature matrix with 5000 max features and most 3 consecutive words range.

4.Train test split whith 33% test size

5.Train seven different ML algorithms to the processed dataset.

TF-IDF:

1.Take train(20800 data) and test(5200) data from kaggle

2.Preprocessing: Make new column using News Title and Whole News and News Author

3.Use TF-IDF transformer to transfer the train and test data into feature matrix.

4.Default train test split

5.Train six different ML algorithms to the processed dataset.

LSTM:

A sequential deep learning model has been implemented using LSTM architecture for binary text classification that performed better with around 99% accuracy. The dataset has been collected from Kaggle and is of the size 20800. The task was to predict if the news is fake or real. Therefore, the pretrained Glove text embedding algorithm has been used as a text vectorization technique. Besides, several classical

models have been implemented with BOW, TF-IDF text vectorization methods. Therefore, the LSTM based deep learning model performs better to classify news.

Stage 5: Test—Try Your Solutions Out

Test: the fifth and final phase of the design thinking process, where you test solutions to derive a deep understanding of the product and its users.

Best One Of ML model Algorithms

Here we have build all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. We have used Logistic Regression, Stochastic gradient descent,Random forest, GBC, xgboost, DecisionTree, Multinomial Naive Baye and Bernoulli Naive Baye classifiers . Each of the extracted features were used in all of the classifiers. Once fitting the model, we compared the accuracy score and checked the confusion matrix.

accuracy score

The highest accuracy score we are getting is 87.04 but don't worry the model was trained with 61,000+ recored it will perform well Our finally selected and *best performing classifier was Logistic Regression which was then saved on disk with name model.plk . Once you clone this repository, this model will be copied to your machine and will be used for prediction. It takes an news article as input from user then shown to user whether it is true or Fake. model.plk* is used to deploy the model usinf Flask.