Capstone Project 2 : MileStone Report

Project Title :

“Fake News Classifier”

Project Mentor : Vikas

What is the Problem to be solved?

In Daily routine we will read a number of articles about trends , politics , technology,food etc., but we just read and will come to our opinion and we will not think whether that article is Fake or Real

Who is this project beneficial for?

This project is used for all of us who have a habit of reading news articles. So this project will help us to identify a given article is Fake or Real

So *What is the data used for the project? And where I received the data from?*

So the data consists of all News records about different topics in the world and I have extracted the data from Kaggle website. This is a multivariate data consisting of Id, title, author, text,label, etc.

Techniques that I would use to solve the problem:

1) Data Wrangling - I will observe the data for all the missing values, incorrect entries and based on the observation I will clean the data for further use and apply the Modern NLP techniques to transform text data into vector form

2) Exploratory Data Analysis - I would plot the relation between different parameters in the dataset and based on that I can find the trends for storytelling.

3) Machine Learning Model / Deep Learning - After all the initial work on the dataset it would be time to build a Machine Learning Model to find the given article is Fake or Real i may use the LSTM or a simple Naive Bayes.

What is the methods I will use to deliver the project?

I will be using the following tools to work on the project :-

- Jupyter Notebook for the major part of the project. - I will be using Python version 3.x for writing the code. - I will use Google doc for submission on the in process project reports. - Powerpoint for Data Storytelling presentation.

References:

- Springboard Curriculum material - DataCamp courses - Google, Stackoverflow, Github, etc. - And any other Open Resources.

Data Wrangling Report:

Data Source:

This dataset is downloaded from the Kaggle website which holds many competitions for the Data science learners with real-world problem statements and data sets.

Initial Data Observations:

The data is split into two datasets: train and test. The train dataset has a shape of (18285,5) and the test dataset is of the shape (2671,4) with label column dropped. The features of the train dataset are [‘id’ ,’title’ , ‘author’, ‘text’ ,’label’’]. The column types of the dataset are as follows:

Column Name	**Type**
id	int64
title	object
author	object
text	object
label	int64

Data Wrangling process:

Missing Value Treatment:

I can see multiple missing values for title , author , text since they are text fields so I dropper all the missing values because we can’t impute like for numeric type with mean , median etc.,

Outlier Treatment:

Here we can’t say for the text fields whether they are outliers or not . Outliers can only identified for numeric columns so I Ignored Outlier treatment .

Now the dataset is mostly clean and ready for further process of Exploratory Data Analysis, Data Visualization and Machine Learning modelling.

Data Processing :

Now all the text data should be handled . I am considering the text or title field for identifying given text is Fake or Real because it will not dependent on the type of Id or author based on the content we can say it is Fake or Real .

Here for that I defined a Corpus with all the text field data then I did the following steps :

- Remove all the stopwords because they are no use for our prediction and they will decrease our accuracy of model if we are not removed
- Bring the root words for each word by using Stemming or lemmatizing
- Now convert all the words into lower case and split the sentence and append to the list
- Now Each sentence is list of words with no stop words and with root words now we have to do count vectorizer or tf-idf vector to convert words into vectors become model can’t take text fields .Count vectorizer is Bag-of-model technique and Tf-idf model will assign weights to words based on their importance if the word is repeated in all the sentences then weight will be less and if word is unique to that sentence then it will have more weight

Machine Learning:

The general machine learning framework is outlined below:

1. **Prediction Engineering:** State the business need, translate into a machine learning problem, and generate labeled examples from a dataset.
2. **Feature Engineering:** Extract predictor variables — features — from the raw data for each of the labels.
3. **Modeling:** Train a machine learning model on the features, tune for the business need, and validate predictions before deploying to new data.

Type of Machine Learning:

The dataset comes under Supervised Learning since it has labelled data. Basically supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data. Supervised learning classified into two categories of algorithms:

● **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.

● **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. In this case our Supervised Learning is a Regression algorithm where we have to predict the Prices of the Airline tickets.

The Process of Prediction Engineering:

Prediction engineering requires guidance both from the business viewpoint to figure out the right problem to solve as well as from the data scientist to determine how to translate the business needs into a machine learning problem. The inputs to prediction engineering are the *parameters* which define the prediction problem for the business requirement , and the historical dataset for finding examples of what we want to predict.

Feature Engineering:

Feature engineering, the second step in the machine learning pipeline, takes in the label times from the first step — prediction engineering — and a raw dataset that needs to be refined. Feature engineering means building features for each label while filtering the data used for the feature based on the label’s cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

The Machine Learning Modeling Process:

The outputs of prediction and feature engineering are a set of label times, historical examples of what we want to predict, and features, predictor variables used to train a model to predict the label. The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business needs, and validating it on holdout data.

Machine Learning Algorithms:

The Algorithms used to build a model and predict the outcomes are:

1. Naive Bayes Classifier: It works well with the text data especially multinomial Naive Bayes because when a new word come in the test data it will give some

prior probability instead of 0 so it makes Multinomial NB different from others.

2. Long Short Term memory (LSTM): **Long short-term memory** (**LSTM**) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).