

**Fake -News Detection Project**

Submitted by:

ARUMOY SAHA

**ACKNOWLEDGMENT**

**INTRODUCTION**

* **Problem statement:**
* The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.
* In this project, you are given a dataset in the fake-news\_data.zip folder. The folder contains a CSV files train\_news.csv and you have to use the train\_news.csv data to build a model to predict whether a news is fake or not fake. You have to try out different models on the dataset, evaluate their performance, and finally report the best model you got on the data and its performance.
* Conceptual Background of the Domain Problem:

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that Fake news across multiple platforms, there is a lack of models for detection of fake news

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

* Review of Literature: Now a days balancing an environment on social media platform is extremely important. Describing as “Fake or Real” cyber bullying puts targets under attack from a barrage.
* Motivation for the Problem Undertaken:

This project helps me understand the toxic comments classification problem in Fake news detection problem. With the right set of datasets in hand I have built a model that helps the enterprise take the right decision that is whether to focus on a fake or real set of news headlines. This also motivate learn about text classification problem in fake news across channels in details. This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes many categories of comments, we can do good amount of data exploration and derive some interesting features using the news column available.

We built a model that can differentiate between news and its categories.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem:

Our goal is to build a prototype of detection of fake news classifier which can used to classify news so that it can be controlled and restricted from spreading across media.

. I have used various classifier model to classify the comments in terms fake or real news and also used cross validation to remove overfitting problem while predicted the correct outcome and validate the model.

* Data sources are provided internally by the enterprise.
* There are 6 columns in the dataset provided to you. The description of each of the column is given below:
* “id”: Unique id of each news article
* “headline”: It is the title of the news.
* “news”: It contains the full text of the news article
* “Unnamed:0”: It is a serial number
* “written\_by”: It represents the author of the news article
* “label”: It tells whether the news is fake (1) or not fake (0).
* Data Pre-processing:

In the data pre-processing stage, I have found out if there is any missing data in dataset, for a particular column if there are any outliers present and how to handle the outliers. I have also found the total shape of the data set. I have also found out the dataset description using describe method. So, in this pre-processing process I have mainly cleansed the data and prepared the right set of data for further processing & for predicting the model.

* Data Inputs- Logic- Output Relationships:

To find out the relationship between all the input variable I have used correlation function and find out whether there is a positive/negative relationship between a pair of variables. From this describe function that also known as Five-point summary analysis if there are any outliers are present for a particular column. Also five point summary analysis was done for the target variable to explore & understand the data in a better way.

* State the set of assumptions (if any) related to the problem under consideration:

Since all the dataset provided and defined properly so in this dataset, I assume label as the target variable for this project. Rest of the parameters are used as input variables.

* Hardware and Software Requirements and Tools Used:

For this particular dataset the Hardware is used Windows as operating system, and the software used are mainly Jupyter notebook for model building and various internal packages that are defined in the anaconda/jupyter notebook.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods):

For this particular project I have used different classification models to predict the outcome of this dataset. After the model implementation PassiveAggressiveClassifier method predicted the best outcome out of all the models in terms of accuracy score and also I have used cross validation to flag the problem related overfitting or selection bias for the dataset and hence we can use this model for further evaluation.

* Testing of Identified Approaches (Algorithms):

I have used mainly different classification methods to get the outcome of the house price prediction and 67% data used for training purpose and rest 33% are used for testing the prediction of the accuracy score for this machine learning model building process.

* Run and Evaluate selected models: I have used various classification models for this dataset.

**Out of all the machine learning models used I have selected PassiveAgressiveClassifier model for further evaluation of this project.**

* Key Metrics for success in solving problem under consideration

The key metrics that were mainly taken into consideration were the followings:

These are all the prime metrics under consideration.

“headline”: It is the title of the news.

“news”: It contains the full text of the news article

“written\_by”: It represents the author of the news article

“label”: It tells whether the news is fake (1) or not fake (0).

* Interpretation of the Results:
* **PassiveAgressiveClassifier** best result for this dataset.
* I have also found out the confusion matrix and classification report for the classification algorithm that is used for this dataset.

**CONCLUSION**

* Key Findings and Conclusions of the Study:
* I used various classification methods and out of all machine learning algorithm used, PassiveAgressiveClassifier yields the best results.
* Learning Outcomes of the Study in respect of Data Science:

As per as learning outcomes is concerned, I have learnt the following things in this project:

* Algorithm need to be used by understanding the dataset for the classification model.
* From describe method we can get some knowledge related to outliers present in the particular columns (large difference between 75th percentile and maximum percentile)
* I also understand the visualization of related features and importance related to dataset.

Challenges:

* It was difficult to load the dataset in notebook as it took some time.
* Running each line code was a bit slow in notebook, possibly due to low CPU configuration.
* Limitations of this work and Scope for Future Work:
* Since I have only used a sample dataset, hence sometimes it is difficult to understand the overall impact of this project while filter out the toxic comments in public forum.