Submitted by:

Joseph davis

INTRODUCTION

Fake news is false or misleading content presented as news and communicated in formats spanning spoken, written, printed, electronic, and digital communication.It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue.Fake news spread through social media has become a serious problem, with the potential of it resulting in mob violence, suicides etc as a result of misinformation circulated on social media.

DATA OVERVIEW

This dataset consists of about 45000 articles consisting of fake as well as real news. Our aim is train our model so that it can correctly predict whether a given piece of news is real or fake.The fake and real news data is given in two separate datasets with fake and real news contains around 23000,22000 articles respectively

SPECIFICATION

Title	Headline of a news article
Text	The content of article
Subject	The category which news belongs
Date	The date at which the article was posted
Target	A label that seprates true news and fake news

GOAL

To buid a predictive model and find out the news is real or fake

APPROACH

Hence, this analysis will be divided into five stages:

1)Exploratory data analysis (EDA)
2)Data cleaning(NLP algorithm)
3)Modeling
4)Hyperparameter tuning

EXPLORATORY DATA ANALYTICS(EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.Hence, the goal for this section is to take a glimpse on the data as well as any irregularities so that we can correct on the next section,ie Data cleaning

DATA CLEANING(NLP)

First we are seperating the content data from dataset for an ease of processing.

Following are the different steps that i have followed

1)Tokenisation-Tokenization is a common task in Natural Language Processing (NLP).Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords.

2)Removing Punctuations-Punctuation is the use of spacing, conventional signs and certain typographical devices as aids to the understanding and correct reading of written textPunctuation does not add much more meaning in modelling.So we can remove it by using regular expressions

3)Stemming-It is the process of reducing the word to its word stem that affixes to suffixes and prefixes or to roots of words known as a lemma. In simple words stemming is reducing a word to its base word or stem in such a way that the words of similar kind lie under a common stem.Snowball stemmer is a stemming algorithm that iam using here

4)Removing stopwords-Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc.In computing, stop words are words that are filtered out before or after the natural language data (text) are processed

5)TF IDF vectorization(term frequency inverse document- frequency)

when doing natural language processing, words must be converted into vectors that machine learning algorithms can make use of.In information retrieval, tf idf or TFIDF, short for term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling.

MODELS USED

These are the models that iam used for classification

1. Logistic Regression
2. Random Forest Classifier
3. Xgb classifier
4.Support vector machine
5.KNN classification
6.Ridge classifier

HYPER PARAMETER TUNING

We import GridsearchCV for parameter tuning and found out the best parameters that fits our model

RESULT

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

True Positives : The cases in which we predicted YES and the actual output was also YES.(TP)

True Negatives : The cases in which we predicted NO and the actual output was NO.(TN)

False Positives : The cases in which we predicted YES and the actual output was NO.(FP)

False Negatives : The cases in which we predicted NO and the actual output was YES.(FN)

Accuracy is the ratio of corrected predictions to total number of input samples

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Logistic regression-99.25 Ridge classifier-98.82
Random forest classifier-99.09
XG boost-99.79
SVM-99.2
KNN-72.56

XG boost is our best model