Introduction:

Explain the properties of the chosen data set and what you will be doing with it:

There are two data sets, one of them contains all the fake news, and the other contains all the true or verified news. Using these data sets, we will create 3 working machine learning models for data classification between fake and true news, with the help of natural language processing (NLP).

Mention the two (or more) machine learning techniques that you will be using:

We will be using 3 machine learning models:

Logistic Regression Support Vector Machine Naive Bayes

Background:

Describe the mechanics of the selected machine learning techniques:

Logistic Regression:

Logistic regression is a simple yet very effective classification algorithm so it is commonly used for many binary classification tasks.Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y).

Support Vector Machine: Classification

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

Naive Bayes:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Describe what rescaling and normalisation are and why they are important:

Normalization is a technique for organizing data in a database. It is important that a database is normalized to minimize redundancy (duplicate data) and to ensure only related data is stored in each table. It also prevents any issues stemming from database modifications such as insertions, deletions, and updates.

Describe what cross validation is:

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. It is a resampling technique with the fundamental idea of splitting the dataset into 2 parts- training data and test data. Train data is used to train the model and the unseen test data is used for prediction. If the model performs well over the test data and gives good accuracy, it means the model hasn’t overfitted the training data and can be used for prediction. Here we split our data into K parts, let’s use K=3 for a toy example. If we have 3000 instances in our dataset, We split it into three parts, part 1, part 2 and part 3. We then build three different models, each model is trained on two parts and tested on the third. Our first model is trained on part 1 and 2 and tested on part 3. Our second model is trained to on part 1 and part 3 and tested on part 2 and so on.

Describe what dimensionality reduction and feature selection methods are:

Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimension.

Explain the quantitative measurements that you will be using to quantify the results

We will be using accuracy score as our measurement. Accuracy score is:

Accuracy Score: Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy = [Number of correct predictions / Total number of predictions]

In [1]:

```
# Importing necessary libraries.
```

In [2]:

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

In [3]:

```
# Importing the fake dataset.
fake = pd.read_csv('Fake.csv', nrows = 5000)
fake
```

Out[3]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	It s no secret Republicans are salivating to f...	News	August 18, 2016
4996	Justice Department Announces It Will No Longe...	Republicans are about to lose a huge source of...	News	August 18, 2016
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa...	A pawn working for Donald Trump claimed that w...	News	August 18, 2016
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	Fox News is desperate to sabotage Hillary Clin...	News	August 18, 2016
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa...	As Donald Trump s campaign continues to sink d...	News	August 18, 2016

5000 rows × 4 columns

In [4]:

```
# Importing the true dataset.
true = pd.read_csv('True.csv', nrows = 5000)
true
```

Out[4]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
...	...	...	...	...
4995	U.S. Agriculture secretary nominee submits eth...	(Reuters) - U.S. President Donald Trump's nomi...	politicsNews	March 13, 2017
4996	Trump aides attack agency that will analyze he...	WASHINGTON (Reuters) - Aides to U.S. President...	politicsNews	March 12, 2017
4997	Highlights: The Trump presidency on March 12 a...	(Reuters) - Highlights of the day for U.S. Pre...	politicsNews	March 12, 2017
4998	Obama lawyers move fast to join fight against ...	WASHINGTON (Reuters) - When Johnathan Smith re...	politicsNews	March 13, 2017
4999	Mike Pence to tour Asia next month amid securi...	JAKARTA (Reuters) - U.S. Vice President Mike P...	politicsNews	March 13, 2017

5000 rows × 4 columns

In [5]:

```
# Concatinating title and text column for the final datasets.
```

In [6]:

```
fake['Description'] = fake['title'] + " " + fake['text']
fake['Description']
#fake['Description'][0]
```

Out[6]:

```
0       Donald Trump Sends Out Embarrassing New Year’...
1       Drunk Bragging Trump Staffer Started Russian ...
2       Sheriff David Clarke Becomes An Internet Joke...
3       Trump Is So Obsessed He Even Has Obama’s Name...
4       Pope Francis Just Called Out Donald Trump Dur...
                              ...
4995    FBI Warns Republicans: Do Not Leak Clinton Em...
4996    Justice Department Announces It Will No Longe...
4997    WATCH: S.E. Cupp Destroys Trump Adviser’s ‘Fa...
4998    WATCH: Fox Hosts Claim Hillary Has Brain Dama...
4999    CNN Panelist LAUGHS In Corey Lewandowski’s Fa...
Name: Description, Length: 5000, dtype: object
```

In [7]:

```
true['Description'] = true['title'] + " " + true['text']
true['Description']
#true['Description'][0]
```

Out[7]:

```
0       As U.S. budget fight looms, Republicans flip t...
1       U.S. military to accept transgender recruits o...
2       Senior U.S. Republican senator: 'Let Mr. Muell...
3       FBI Russia probe helped by Australian diplomat...
4       Trump wants Postal Service to charge 'much mor...
                              ...
4995    U.S. Agriculture secretary nominee submits eth...
4996    Trump aides attack agency that will analyze he...
4997    Highlights: The Trump presidency on March 12 a...
4998    Obama lawyers move fast to join fight against ...
4999    Mike Pence to tour Asia next month amid securi...
Name: Description, Length: 5000, dtype: object
```

In [8]:

```
# Adding label: 1, for all the true descriptions.
```

In [9]:

```
true['label'] = 1
true
```

Out[9]:

	title	text	subject	date	Description	label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	As U.S. budget fight looms, Republicans flip t...	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	U.S. military to accept transgender recruits o...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	Senior U.S. Republican senator: 'Let Mr. Muell...	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	FBI Russia probe helped by Australian diplomat...	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	Trump wants Postal Service to charge 'much mor...	1
...	...	...	...	...	...	...
4995	U.S. Agriculture secretary nominee submits eth...	(Reuters) - U.S. President Donald Trump's nomi...	politicsNews	March 13, 2017	U.S. Agriculture secretary nominee submits eth...	1
4996	Trump aides attack agency that will analyze he...	WASHINGTON (Reuters) - Aides to U.S. President...	politicsNews	March 12, 2017	Trump aides attack agency that will analyze he...	1
4997	Highlights: The Trump presidency on March 12 a...	(Reuters) - Highlights of the day for U.S. Pre...	politicsNews	March 12, 2017	Highlights: The Trump presidency on March 12 a...	1
4998	Obama lawyers move fast to join fight against ...	WASHINGTON (Reuters) - When Johnathan Smith re...	politicsNews	March 13, 2017	Obama lawyers move fast to join fight against ...	1
4999	Mike Pence to tour Asia next month amid securi...	JAKARTA (Reuters) - U.S. Vice President Mike P...	politicsNews	March 13, 2017	Mike Pence to tour Asia next month amid securi...	1

5000 rows × 6 columns

In [10]:

```
# Adding label: 0, for all the fake descriptions.
```

In [11]:

```
fake['label'] = 0
fake
```

Out[11]:

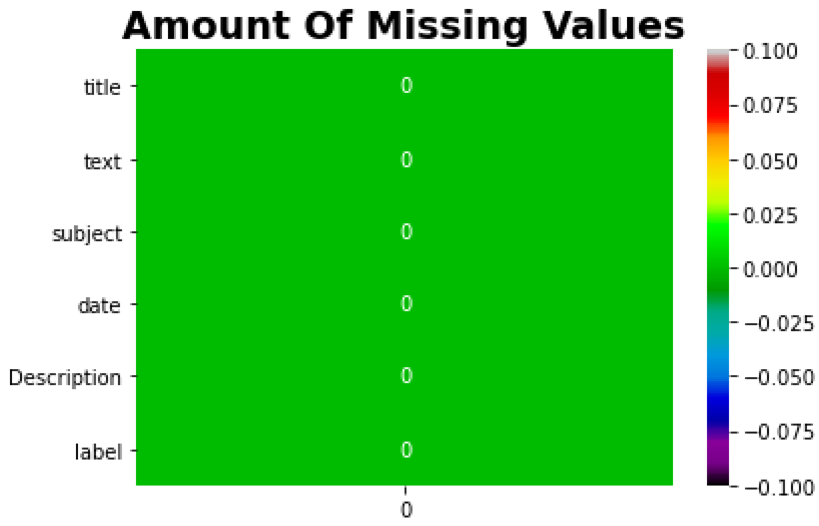
	title	text	subject	date	Description	label
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	Donald Trump Sends Out Embarrassing New Year’...	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	Drunk Bragging Trump Staffer Started Russian ...	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	Sheriff David Clarke Becomes An Internet Joke...	0
3	Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	Trump Is So Obsessed He Even Has Obama’s Name...	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	Pope Francis Just Called Out Donald Trump Dur...	0
...	...	...	...	...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	It s no secret Republicans are salivating to f...	News	August 18, 2016	FBI Warns Republicans: Do Not Leak Clinton Em...	0
4996	Justice Department Announces It Will No Longe...	Republicans are about to lose a huge source of...	News	August 18, 2016	Justice Department Announces It Will No Longe...	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser’s ‘Fa...	A pawn working for Donald Trump claimed that w...	News	August 18, 2016	WATCH: S.E. Cupp Destroys Trump Adviser’s ‘Fa...	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	Fox News is desperate to sabotage Hillary Clin...	News	August 18, 2016	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	0
4999	CNN Panelist LAUGHS In Corey Lewandowski’s Fa...	As Donald Trump s campaign continues to sink d...	News	August 18, 2016	CNN Panelist LAUGHS In Corey Lewandowski’s Fa...	0

5000 rows × 6 columns

Checking for any missing values on both the datasets.

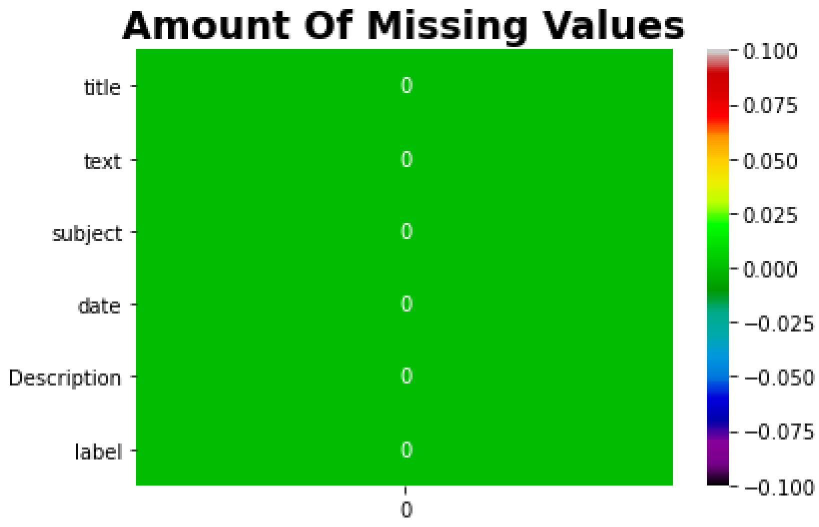
In [12]:

```
plt.title('Amount Of Missing Values',fontsize=19,fontweight='bold')
sns.heatmap(fake.isna().sum().to_frame(),annot=True,cmap='nipy_spectral')
plt.show()
```



In [13]:

```
plt.title('Amount Of Missing Values',fontsize=19,fontweight='bold')
sns.heatmap(true.isna().sum().to_frame(),annot=True,cmap='nipy_spectral')
plt.show()
```



Creating a new data frame only with the required columns.

In [14]:

```
true_final = true[['Description', 'label']]
true_final
```

Out[14]:

	Description	label
0	As U.S. budget fight looms, Republicans flip t...	1
1	U.S. military to accept transgender recruits o...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	1
3	FBI Russia probe helped by Australian diplomat...	1
4	Trump wants Postal Service to charge 'much mor...	1
...	...	...
4995	U.S. Agriculture secretary nominee submits eth...	1
4996	Trump aides attack agency that will analyze he...	1
4997	Highlights: The Trump presidency on March 12 a...	1
4998	Obama lawyers move fast to join fight against ...	1
4999	Mike Pence to tour Asia next month amid securi...	1

5000 rows × 2 columns

In [15]:

```
fake_final = fake[['Description', 'label']]
fake_final
```

Out[15]:

	Description	label
0	Donald Trump Sends Out Embarrassing New Year'...	0
1	Drunk Bragging Trump Staffer Started Russian ...	0
2	Sheriff David Clarke Becomes An Internet Joke...	0
3	Trump Is So Obsessed He Even Has Obama's Name...	0
4	Pope Francis Just Called Out Donald Trump Dur...	0
...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	0
4996	Justice Department Announces It Will No Longe...	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa...	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	0
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa...	0

5000 rows × 2 columns

Joining both the datasets row wise, and creating the final dataset.

In [16]:

```
df = pd.concat([true_final, fake_final])
df.head(5)
```

Out[16]:

	Description	label
0	As U.S. budget fight looms, Republicans flip t...	1
1	U.S. military to accept transgender recruits o...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	1
3	FBI Russia probe helped by Australian diplomat...	1
4	Trump wants Postal Service to charge 'much mor...	1

In [17]:

```
df.tail(5)
```

Out[17]:

	Description	label
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	0
4996	Justice Department Announces It Will No Longe...	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser’s ‘Fa...	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	0
4999	CNN Panelist LAUGHS In Corey Lewandowski’s Fa...	0

In [18]:

```
df
```

Out[18]:

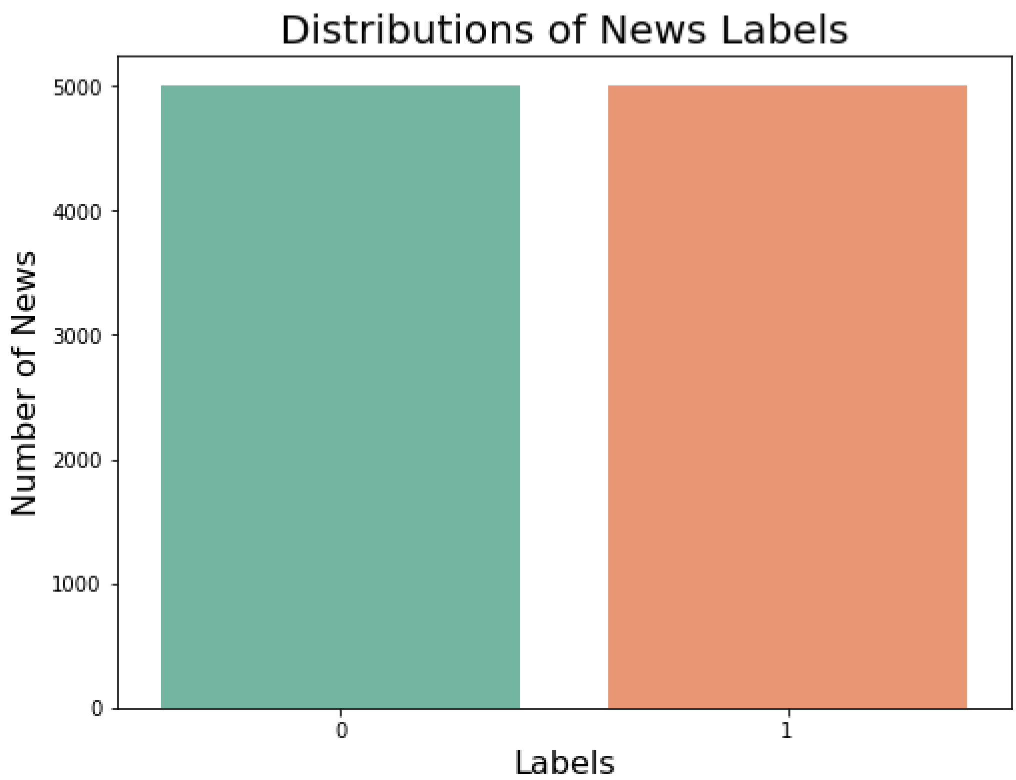
	Description	label
0	As U.S. budget fight looms, Republicans flip t...	1
1	U.S. military to accept transgender recruits o...	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	1
3	FBI Russia probe helped by Australian diplomat...	1
4	Trump wants Postal Service to charge 'much mor...	1
...	...	...
4995	FBI Warns Republicans: Do Not Leak Clinton Em...	0
4996	Justice Department Announces It Will No Longe...	0
4997	WATCH: S.E. Cupp Destroys Trump Adviser's 'Fa...	0
4998	WATCH: Fox Hosts Claim Hillary Has Brain Dama...	0
4999	CNN Panelist LAUGHS In Corey Lewandowski's Fa...	0

10000 rows × 2 columns

In [19]:

```
# Final labels Countplot

plt.figure(figsize=(8,6))
sns.countplot(df.label, palette='Set2')
plt.title('Distributions of News Labels',fontsize=20)
plt.xlabel('Labels', fontsize=16)
plt.ylabel('Number of News', fontsize=16)
plt.show()
```



Cleaning the dataset:

https://medium.com/analytics-vidhya/data-cleaning-in-natural-language-processing-1f77ec1f6406 (https://medium.com/analytics-vidhya/data-cleaning-in-natural-language-processing-1f77ec1f6406)

- Removing Puntuations
- Stopword Removal
- Lemmatizing the cleaned description column
- Normalizing the data
- Using principal components to extract the features
- Creating X and y training and testing sets

In [20]:

```
# Importing necessarry libraries to clean the data set
import nltk
import string
import re
```

In [21]:

```
# Removing Puntuation:

def remove_punct(text):
    text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub('[0-9]+', '', text)
    return text
```

In [22]:

```
# Stopword Removal:

from nltk.corpus import stopwords
", ".join(stopwords.words('english'))
STOPWORDS = set(stopwords.words('english'))

def remove_stopwords(text):
    """custom function to remove the stopwords"""
    return " ".join([word for word in str(text).split() if word not in STOPWORDS])
```

In [23]:

```
# Applying all the defined functions in the data

df['Description'] = df['Description'].apply(lambda x: remove_punct(x))
df['clean_Description'] = df['Description'].apply(lambda text: remove_stopwords(text))
df.head(5)
```

Out[23]:

	Description	label	clean_Description
0	As US budget fight looms Republicans flip thei...	1	As US budget fight looms Republicans flip fisc...
1	US military to accept transgender recruits on ...	1	US military accept transgender recruits Monday...
2	Senior US Republican senator Let Mr Mueller do...	1	Senior US Republican senator Let Mr Mueller jo...
3	FBI Russia probe helped by Australian diplomat...	1	FBI Russia probe helped Australian diplomat ti...
4	Trump wants Postal Service to charge much more...	1	Trump wants Postal Service charge much Amazon ...

In [24]:

```
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Karsten\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\Karsten\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[24]:

```
True
```

In [25]:

```
# Lemmatizing the cleaned description column
# Stemming uses the stem of the word, while lemmatization uses the context in which the
# Lemmatization carries out a morphological analysis of the words, the chatbot is able

from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
wordnet_map = {"N":wordnet.NOUN, "V":wordnet.VERB, "J":wordnet.ADJ, "R":wordnet.ADV}

def lemmatize_words(text):
    pos_tagged_text = nltk.pos_tag(text.split())
    return " ".join([lemmatizer.lemmatize(word, wordnet_map.get(pos[0], wordnet.NOUN))

df["Description_lemmatized"] = df["clean_Description"].apply(lambda text: lemmatize_wor
df.head(5)
```

Out[25]:

	Description	label	clean_Description	Description_lemmatized
0	As US budget fight looms Republicans flip thei...	1	As US budget fight looms Republicans flip fisc...	As US budget fight loom Republicans flip fisca...
1	US military to accept transgender recruits on ...	1	US military accept transgender recruits Monday...	US military accept transgender recruit Monday ...
2	Senior US Republican senator Let Mr Mueller do...	1	Senior US Republican senator Let Mr Mueller jo...	Senior US Republican senator Let Mr Mueller jo...
3	FBI Russia probe helped by Australian diplomat...	1	FBI Russia probe helped Australian diplomat ti...	FBI Russia probe help Australian diplomat tipo...
4	Trump wants Postal Service to charge much more...	1	Trump wants Postal Service charge much Amazon ...	Trump want Postal Service charge much Amazon s...

In [26]:

```
1 df.head(20)
```

Out[26]:

	Description	label	clean_Description	Description_lemmatized
0	As US budget fight looms Republicans flip thei...	1	As US budget fight looms Republicans flip fisc...	As US budget fight loom Republicans flip fisca...
1	US military to accept transgender recruits on ...	1	US military accept transgender recruits Monday...	US military accept transgender recruit Monday ...
2	Senior US Republican senator Let Mr Mueller do...	1	Senior US Republican senator Let Mr Mueller jo...	Senior US Republican senator Let Mr Mueller jo...
3	FBI Russia probe helped by Australian diplomat...	1	FBI Russia probe helped Australian diplomat ti...	FBI Russia probe help Australian diplomat tipo...
4	Trump wants Postal Service to charge much more...	1	Trump wants Postal Service charge much Amazon ...	Trump want Postal Service charge much Amazon s...
5	White House Congress prepare for talks on spen...	1	White House Congress prepare talks spending im...	White House Congress prepare talk spend immigr...
6	Trump says Russia probe will be fair but timel...	1	Trump says Russia probe fair timeline unclear ...	Trump say Russia probe fair timeline unclear N...
7	Factbox Trump on Twitter Dec Approval rating...	1	Factbox Trump Twitter Dec Approval rating Amaz...	Factbox Trump Twitter Dec Approval rating Amaz...
8	Trump on Twitter Dec Global Warming The foll...	1	Trump Twitter Dec Global Warming The following...	Trump Twitter Dec Global Warming The following...
9	Alabama official to certify Senatorelect Jones...	1	Alabama official certify Senatorelect Jones to...	Alabama official certify Senatorelect Jones to...
10	Jones certified US Senate winner despite Moore...	1	Jones certified US Senate winner despite Moore...	Jones certify US Senate winner despite Moore c...
11	New York governor questions the constitutional...	1	New York governor questions constitutionality ...	New York governor question constitutionality f...
12	Factbox Trump on Twitter Dec Vanity Fair Hil...	1	Factbox Trump Twitter Dec Vanity Fair Hillary ...	Factbox Trump Twitter Dec Vanity Fair Hillary ...
13	Trump on Twitter Dec Trump Iraq Syria The fo...	1	Trump Twitter Dec Trump Iraq Syria The followi...	Trump Twitter Dec Trump Iraq Syria The followi...
14	Man says he delivered manure to Mnuchin to pro...	1	Man says delivered manure Mnuchin protest new ...	Man say deliver manure Mnuchin protest new US ...
15	Virginia officials postpone lottery drawing to...	1	Virginia officials postpone lottery drawing de...	Virginia official postpone lottery draw decide...
16	US lawmakers question businessman at Trump To...	1	US lawmakers question businessman Trump Tower ...	US lawmaker question businessman Trump Tower m...
17	Trump on Twitter Dec Hillary Clinton Tax Cut...	1	Trump Twitter Dec Hillary Clinton Tax Cut Bill...	Trump Twitter Dec Hillary Clinton Tax Cut Bill...
18	US appeals court rejects challenge to Trump vo...	1	US appeals court rejects challenge Trump voter...	US appeal court reject challenge Trump voter f...
19	Treasury Secretary Mnuchin was sent giftwrappe...	1	Treasury Secretary Mnuchin sent giftwrapped bo...	Treasury Secretary Mnuchin send giftwrapped bo...

In [27]:

```
!pip install transformers
```

```
Requirement already satisfied: transformers in c:\users\karsten\anaconda3\li
b\site-packages (4.14.1)
Requirement already satisfied: pyyaml>=5.1 in c:\users\karsten\anaconda3\lib
\site-packages (from transformers) (5.3.1)
Requirement already satisfied: sacremoses in c:\users\karsten\anaconda3\lib
\site-packages (from transformers) (0.0.46)
Requirement already satisfied: filelock in c:\users\karsten\anaconda3\lib\si
te-packages (from transformers) (3.0.12)
Requirement already satisfied: packaging>=20.0 in c:\users\karsten\anaconda3
\lib\site-packages (from transformers) (21.3)
Requirement already satisfied: regex!=2019.12.17 in c:\users\karsten\anacond
a3\lib\site-packages (from transformers) (2020.6.8)
Requirement already satisfied: huggingface-hub<1.0,>=0.1.0 in c:\users\karst
en\anaconda3\lib\site-packages (from transformers) (0.2.1)
Requirement already satisfied: requests in c:\users\karsten\anaconda3\lib\si
te-packages (from transformers) (2.24.0)
Requirement already satisfied: tqdm>=4.27 in c:\users\karsten\anaconda3\lib
\site-packages (from transformers) (4.47.0)
Requirement already satisfied: numpy>=1.17 in c:\users\karsten\anaconda3\lib
\site-packages (from transformers) (1.18.5)
Requirement already satisfied: tokenizers<0.11,>=0.10.1 in c:\users\karsten
\anaconda3\lib\site-packages (from transformers) (0.10.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\karste
n\anaconda3\lib\site-packages (from huggingface-hub<1.0,>=0.1.0->transformer
s) (4.0.1)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\karsten
\anaconda3\lib\site-packages (from packaging>=20.0->transformers) (2.4.7)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\karsten\anacond
a3\lib\site-packages (from requests->transformers) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in c:\users\karsten\anaconda3\li
b\site-packages (from requests->transformers) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\karsten\anacon
da3\lib\site-packages (from requests->transformers) (2020.6.20)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
c:\users\karsten\anaconda3\lib\site-packages (from requests->transformers)
(1.25.9)
Requirement already satisfied: click in c:\users\karsten\anaconda3\lib\site-
packages (from sacremoses->transformers) (7.1.2)
Requirement already satisfied: six in c:\users\karsten\anaconda3\lib\site-pa
ckages (from sacremoses->transformers) (1.15.0)
Requirement already satisfied: joblib in c:\users\karsten\anaconda3\lib\site
-packages (from sacremoses->transformers) (0.16.0)
```

https://www.youtube.com/watch?v=VFp38yj8h3A (https://www.youtube.com/watch?v=VFp38yj8h3A)

In natural language processing most of the data that we handle consists of raw text however machine learning models cannot read or understand text in its raw form they can only work with numbers so the tokenizer's objective will be to translate the text into numbers there are several possible approaches to this conversion and the objective is to find the most meaningful representation we'll take a look at three distinct tokenization algorithms we compare them one to one:

- Word-based tokenizers: https://www.youtube.com/watch?v=nhJxYji1aho (https://www.youtube.com/watch?v=nhJxYji1aho)
- Character-based tokenizers: https://youtu.be/ssLq_EK2jLE (https://youtu.be/ssLq_EK2jLE)

- Subword-based tokenizers: https://youtu.be/zHvTiHr506c (https://youtu.be/zHvTiHr506c) <- We Use as more accuracte,it produces more meaningful subwords

https://www.youtube.com/watch?v=xI0HHN5XKDo (https://www.youtube.com/watch?v=xI0HHN5XKDo)

BERT was trained using the WordPiece tokenization. It means that a word can be broken down into more than one sub-words. To tokenize our texts. BERT was trained using the WordPiece tokenization. It means that a word can be broken down into more than one sub-words. This kind of tokenization is beneficial when dealing with out of vocabulary words, and it may help better represent complicated words. The sub-words are constructed during the training time and depend on the corpus the model was trained on. We could use any other tokenization technique of course, but we'll get the best results if we tokenize with the same tokenizer the BERT model was trained on.

https://www.analyticsvidhya.com/blog/2021/09/an-explanatory-guide-to-bert-tokenizer/ (https://www.analyticsvidhya.com/blog/2021/09/an-explanatory-guide-to-bert-tokenizer/)

In [28]:

```
import numpy as np
from transformers import BertTokenizer
tokenizer=BertTokenizer.from_pretrained('bert-base-uncased')

def sen_to_vec(sentence):
    tokens=tokenizer.tokenize(sentence)
    tokens = ['[CLS]'] + tokens + ['[SEP]']
    T=1519
    padded_tokens=tokens +['[PAD]' for _ in range(T-len(tokens))]
    attn_mask=[ 1 if token != '[PAD]' else 0 for token in padded_tokens  ]
    seg_ids=[0 for _ in range(len(padded_tokens))]
    sent_ids=tokenizer.convert_tokens_to_ids(padded_tokens)
    return np.array(sent_ids)
```

```
None of PyTorch, TensorFlow >= 2.0, or Flax have been found. Models won't be
available and only tokenizers, configuration and file/data utilities can be
used.
```

In [29]:

```
df["Array"]=df["Description_lemmatized"].apply(sen_to_vec)
df.head()
```

Out[29]:

	Description	label	clean_Description	Description_lemmatized	Array
0	As US budget fight looms Republicans flip thei...	1	As US budget fight looms Republicans flip fisc...	As US budget fight loom Republicans flip fisca...	[101, 2004, 2149, 5166, 2954, 8840, 5358, 1064...
1	US military to accept transgender recruits on ...	1	US military accept transgender recruits Monday...	US military accept transgender recruit Monday ...	[101, 2149, 2510, 5138, 16824, 13024, 6928, 20...
2	Senior US Republican senator Let Mr Mueller do...	1	Senior US Republican senator Let Mr Mueller jo...	Senior US Republican senator Let Mr Mueller jo...	[101, 3026, 2149, 3951, 5205, 2292, 2720, 2677...
3	FBI Russia probe helped by Australian diplomat...	1	FBI Russia probe helped Australian diplomat ti...	FBI Russia probe help Australian diplomat tipo...	[101, 8495, 3607, 15113, 2393, 2827, 11125, 59...
4	Trump wants Postal Service to charge much more...	1	Trump wants Postal Service charge much Amazon ...	Trump want Postal Service charge much Amazon s...	[101, 8398, 2215, 10690, 2326, 3715, 2172, 973...

In [30]:

```
df_for_model = df[["Array","label"]]
```

In [31]:

```
df_final_for_model=pd.concat([df_for_model.pop('Array').apply(pd.Series), df_for_model
df_final_for_model=df_final_for_model.fillna(0)
df_final_for_model
```

Out[31]:

	0	1	2	3	4	5	6	7	8	9	..
0	101.0	2004.0	2149.0	5166.0	2954.0	8840.0	5358.0	10643.0	11238.0	10807.0	.
1	101.0	2149.0	2510.0	5138.0	16824.0	13024.0	6928.0	20864.0	2899.0	26665.0	.
2	101.0	3026.0	2149.0	3951.0	5205.0	2292.0	2720.0	26774.0	3105.0	2899.0	.
3	101.0	8495.0	3607.0	15113.0	2393.0	2827.0	11125.0	5955.0	7245.0	6396.0	.
4	101.0	8398.0	2215.0	10690.0	2326.0	3715.0	2172.0	9733.0	22613.0	5862.0	.
...	...	...	...	...	...	...	...	...	...	...	.
4995	101.0	8495.0	19428.0	10643.0	2079.0	2025.0	17271.0	7207.0	10373.0	6764.0	.
4996	101.0	3425.0	2533.0	17472.0	2009.0	2097.0	2053.0	2936.0	2224.0	2797.0	.
4997	101.0	3422.0	7367.0	2452.0	2361.0	20735.0	8398.0	11747.0	1521.0	1055.0	.
4998	101.0	3422.0	4419.0	6184.0	4366.0	18520.0	2038.0	4167.0	4053.0	2138.0	.
4999	101.0	13229.0	5997.0	2923.0	11680.0	1999.0	18132.0	24992.0	28574.0	10344.0	.

10000 rows × 3354 columns

In [32]:

```
# Creating X and y variable
X = df_final_for_model.drop('label', axis=1)
y = df_final_for_model.label
```

In [33]:

```
X
```

Out[33]:

	0	1	2	3	4	5	6	7	8	9	...
0	101.0	2004.0	2149.0	5166.0	2954.0	8840.0	5358.0	10643.0	11238.0	10807.0	...
1	101.0	2149.0	2510.0	5138.0	16824.0	13024.0	6928.0	20864.0	2899.0	26665.0	...
2	101.0	3026.0	2149.0	3951.0	5205.0	2292.0	2720.0	26774.0	3105.0	2899.0	...
3	101.0	8495.0	3607.0	15113.0	2393.0	2827.0	11125.0	5955.0	7245.0	6396.0	...
4	101.0	8398.0	2215.0	10690.0	2326.0	3715.0	2172.0	9733.0	22613.0	5862.0	...
...	...	...	...	...	...	...	...	...	...	...	...
4995	101.0	8495.0	19428.0	10643.0	2079.0	2025.0	17271.0	7207.0	10373.0	6764.0	...
4996	101.0	3425.0	2533.0	17472.0	2009.0	2097.0	2053.0	2936.0	2224.0	2797.0	...
4997	101.0	3422.0	7367.0	2452.0	2361.0	20735.0	8398.0	11747.0	1521.0	1055.0	...
4998	101.0	3422.0	4419.0	6184.0	4366.0	18520.0	2038.0	4167.0	4053.0	2138.0	...
4999	101.0	13229.0	5997.0	2923.0	11680.0	1999.0	18132.0	24992.0	28574.0	10344.0	...

10000 rows × 3353 columns

In [34]:

```
y
```

Out[34]:

```
0       1
1       1
2       1
3       1
4       1
       ..
4995    0
4996    0
4997    0
4998    0
4999    0
Name: label, Length: 10000, dtype: int64
```

In [35]:

```
# Normalizing the data

# Initialize a MinMaxScaler and scale the features to between -1 and 1 to normalize the
# The MinMaxScaler transforms features by scaling them to a given range.
# The fit_transform() method fits to the data and then transforms it. We don’t need to
# Scale the features to between -1 and 1

# Scaling is important in the algorithms such as support vector machines (SVM) and k-ne
# between the data points is important.

from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler((0,1))
X = scaler.fit_transform(X)
```

In [36]:

```
# Using principal components to extract the features explaining up to 90% of variance
# Applying Feature Engineering
# Applying PCA
# The code below has .90 for the number of components parameter.
# It means that scikit-Learn choose the minimum number of principal components such tha

from sklearn.decomposition import PCA

pca = PCA(.90)
X_PCA=pca.fit_transform(X)

print(X.shape)
print(X_PCA.shape)
```

```
(10000, 3353)
(10000, 373)
```

In [37]:

```
X_PCA.shape[1]
```

Out[37]:

```
373
```

In [38]:

```
from sklearn.feature_selection import SelectKBest, chi2
selector = SelectKBest(chi2, k=X_PCA.shape[1])
X_kbest = selector.fit_transform(X, y)
```

Scikit Learn

Implementing different kinds of model based on third-party libraries

In [39]:

```
from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_score, co
#from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
#from sklearn.linear_model import SGDClassifier
#from sklearn.ensemble import BaggingClassifier, RandomForestClassifier , AdaBoostClass
import plotly.graph_objects as go
#import xgboost as xgb
from sklearn.metrics import roc_curve, auc
#from sklearn.ensemble import GradientBoostingClassifier
from sklearn.naive_bayes import GaussianNB
```

In [40]:

```
# Creating the function consisting 3 ML techniques
```

In [41]:

```
#Now,split the dataset into training and testing sets keeping 20% of the data for testi
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

X_train_PCA,X_test_PCA,y_train_PCA,y_test_PCA=train_test_split(X_PCA, y, test_size=0.2,
X_train_kbest,X_test_kbest,y_train_kbest,y_test_kbest=train_test_split(X_kbest, y, test
```

In [42]:

```
def func(X_train,y_train,X_test, y_test):
    model_names = ['LogisticRegression', 'SupportVectorMachine', 'NaiveBayes']
    train_scores = []
    test_scores = []

    models = [LogisticRegression,SVC,GaussianNB]

    for model in models:
        mod = model()
        model_fit = mod.fit(X_train, y_train)

        train_scores.append(model_fit.score(X_train, y_train))
        test_scores.append(model_fit.score(X_test, y_test))

    # dictionary of lists
    dd = {'Model': model_names , 'Training_score': train_scores, 'Testing_score': test_

    result = pd.DataFrame(dd)

    return result
```

In [43]:

```
#Thrird party library results
```

In [44]:

```
func(X_train_PCA,y_train_PCA,X_test_PCA, y_test_PCA)
```

Out[44]:

	Model	Training_score	Testing_score
0	LogisticRegression	0.877125	0.8505
1	SupportVectorMachine	0.965000	0.8675
2	NaiveBayes	0.618375	0.6025

In [45]:

```
func(X_train_kbest,y_train_kbest,X_test_kbest, y_test_kbest)
```

Out[45]:

	Model	Training_score	Testing_score
0	LogisticRegression	0.865375	0.8475
1	SupportVectorMachine	0.950500	0.8685
2	NaiveBayes	0.698125	0.6900

In [46]:

```
X_train,X_test,y_train,y_test=train_test_split(X_kbest, y, test_size=0.2, random_state=
```

Cross validation divides the data repeatedly based on k value we give and generates accuracies for them then takes the mean of it. This will help us compare between kfold cross validation models and scikit learn library models

In [48]:

```
from sklearn.model_selection import cross_val_score

models = [LogisticRegression,SVC,GaussianNB]
for model in models:
  #model_names = ['LogisticRegression', 'SupportVectorMachine', 'NaiveBayes']
  cross_validation = cross_val_score(model(), X_train, y_train, cv=5, scoring="accuracy
  print(f"Mean accuracy of {str(model)} is: {np.mean(cross_validation)}")
```

```
Mean accuracy of <class 'sklearn.linear_model._logistic.LogisticRegression'>
is: 0.8401250000000001
Mean accuracy of <class 'sklearn.svm._classes.SVC'> is: 0.8595
Mean accuracy of <class 'sklearn.naive_bayes.GaussianNB'> is: 0.697125
```

Models Based on First Principle

Logistic Regression Manual Implementation

Using our Function to train the Model

In [49]:

```
# to compare our model's accuracy with sklearn model
from sklearn.linear_model import LogisticRegression
# Logistic Regression
class LogitRegression() :
    def __init__( self, learning_rate, iterations ) :
        self.learning_rate = learning_rate
        self.iterations = iterations

    # Function for model training
    def fit( self, X, Y ) :
        # no_of_training_examples, no_of_features
        self.m, self.n = X.shape
        # weight initialization
        self.W = np.zeros( self.n )
        self.b = 0
        self.X = X
        self.Y = Y

        # gradient descent learning

        for i in range( self.iterations ) :
            self.update_weights()
        return self

    # Helper function to update weights in gradient descent

    def update_weights( self ) :
        A = 1 / ( 1 + np.exp( - ( self.X.dot( self.W ) + self.b ) ) )

        # calculate gradients
        tmp = ( A - self.Y.T )
        tmp = np.reshape( tmp, self.m )
        dW = np.dot( self.X.T, tmp ) / self.m
        db = np.sum( tmp ) / self.m

        # update weights
        self.W = self.W - self.learning_rate * dW
        self.b = self.b - self.learning_rate * db

        return self

    # Hypothetical function  h( x )

    def predict( self, X ) :
        Z = 1 / ( 1 + np.exp( - ( X.dot( self.W ) + self.b ) ) )
        Y = np.where( Z > 0.5, 1, 0 )
        return Y
```

In [50]:

```
model = LogitRegression(learning_rate = 0.01, iterations = 1000)
model.fit(X_train, y_train)
```

Out[50]:

```
<__main__.LogitRegression at 0x22dd5ede5b0>
```

In [51]:

```
y_pred = model.predict( X_test )
```

We are creating our own accuracy fuction called 'accurate' to be used for all other models as well:

In [52]:

```
def accurate(y_true, y_pred):
    accuracy = np.sum(y_true == y_pred) / len(y_true)
    return accuracy
```

In [53]:

```
train_predictions = model.predict(X_train)
print("Logit Clasifier Training Accuracy: ",accurate(y_train,train_predictions)*100 )

test_predictions = model.predict(X_test)
print("Logit Clasifier Testing Accuracy: ",accurate(y_test,test_predictions)*100 )
```

```
Logit Clasifier Training Accuracy:  74.8375
Logit Clasifier Testing Accuracy:  74.4
```

Support Vector Machine Manual Implementation:

In [54]:

```
class SVM:

    def __init__(self, learning_rate=0.001, lambda_param=0.01, n_iters=1000):
        self.lr = learning_rate
        self.lambda_param = lambda_param
        self.n_iters = n_iters
        self.w = None
        self.b = None


    def fit(self, X, y):
        n_samples, n_features = X.shape

        y_ = np.where(y == 0, 0, 1)

        self.w = np.zeros(n_features)
        self.b = 0

        for _ in range(self.n_iters):
            for idx, x_i in enumerate(X):
                condition = y_[idx] * (np.dot(x_i, self.w) - self.b) >= 1
                if condition:
                    self.w -= self.lr * (2 * self.lambda_param * self.w)
                else:
                    self.w -= self.lr * (2 * self.lambda_param * self.w - np.dot(x_i, y
                    self.b -= self.lr * y_[idx]


    def predict(self, X):
        approx = np.dot(X, self.w) - self.b
        return np.sign(approx)
```

In [55]:

```
model = SVM()
model.fit(X_train, y_train)
```

In [56]:

```
train_predictions = model.predict(X_train)
print("SVM Clasifier Training Accuracy: ",accurate(y_train,train_predictions)*100 )

test_predictions = model.predict(X_test)
print("SVM Clasifier Testing Accuracy: ",accurate(y_test,test_predictions)*100 )
```

```
SVM Clasifier Training Accuracy:  49.95
SVM Clasifier Testing Accuracy:  50.2
```

Implementing Naive Bayes Algorithm from Scratch:

In [57]:

```
class NaiveBayes:

    def fit(self, X, y): # X is a numpy array, y is a 1D vector
        n_samples, n_features = X.shape # n_samples are Rows, n_features are Columns
        self._classes = np.unique(y) # Unique Classes
        n_classes = len(self._classes)

        # Calculating Mean variace and Prior for each class
        self._mean = np.zeros((n_classes, n_features),dtype=np.float64)
        self._var = np.zeros((n_classes, n_features),dtype=np.float64)
        self._priors = np.zeros(n_classes,dtype=np.float64)


        for idx, c in enumerate(self._classes):
            X_c = X[y==c]
            self._mean[idx, :] = X_c.mean(axis=0)
            self._var[idx, :] = X_c.var(axis=0)
            self._priors[idx] = X_c.shape[0] /float(n_samples)


    def predict(self,X):
        y_pred = [self._predict(x) for x in X] # Using List Comprehension
        return np.array(y_pred)

    def _predict(self,x):
        posteriors = [] # Adding all the values to choose maximum value for posterior p

        # Calculating posterior probability for each class
        for idx, c in enumerate(self._classes):
            prior = np.log(self._priors[idx])
            posterior = np.sum(np.log(self._pdf(idx,x)))
            posterior = prior + posterior
            posteriors.append(posterior)

        # Return class with Highest posterior probability
        return self._classes[np.argmax(posteriors)]


    def _pdf(self, class_idx, x):
        mean = self._mean[class_idx]
        var = self._var[class_idx]
        numerator = np.exp(- (x-mean)**2 / (2 * var))
        denominator = np.sqrt(2 * np.pi * var)
        return numerator / denominator
```

In [58]:

```
nb = NaiveBayes()
nb.fit(X_train, y_train)
```

In [59]:

```
train_predictions = nb.predict(X_train)
print("Naive Bayes Clasifier Training Accuracy: ",accurate(y_train,train_predictions)*1

test_predictions = nb.predict(X_test)
print("Naive Bayes Clasifier Testing Accuracy: ",accurate(y_test,test_predictions)*100
```

```
Naive Bayes Clasifier Training Accuracy:  69.8125
Naive Bayes Clasifier Testing Accuracy:  69.0
```

Experiments:

Describe the steps that you used to process the data set:

To process the data set:

- After importing both the data sets; consists of true and fake news with upto 5000 rows
- We created a text based description column
- Labels for each type of news has been added, where 1 represents true news and 0 represents fake news.
- After checking for any missing values, the final data frame has been created using only the description and label columns.
- With the final data set, we used NLP techniques to clean to data by removing all the stop words and punctuations. Then we used lemmatization and Bert tokenization to get the array form required.
- For the last experiment we checked the cross-validation score, and it didn't provide much help, as accuracies for every model remains similar to that of sklearn methods.

Describe the experiments that you carried out

- first we scale the predictor variables using Minmaxscaler
- Used PCA to get the most important features that explains upto 90% of the variance
- then we used kbest method for feature selection to compare between PCA and kBest methods

Describe the implementation of the 3 ML techniques chosen

- For all three models, that is Logistic, Support Vector and Naive Bayes' we used our own models. The models are build from scratch based on the first principle of python language
- For accuracy we used own accuracy model on all of them to get the result
  - def accurate(y_true, y_pred):
  - accuracy = np.sum(y_true == y_pred) / len(y_true)
  - return accuracy

Compare your implementation of these techniques using the dataset against a third-party implementation of the same techniques.

The comparision of the results:

- For Logistic the training and testing accuracies using third party libraries are: 86% & 84% and using the first principle is: 74.8% & 74.4%.
- For SVM the training and testing accuracies using third party libraries are: 95% & 86% and using the first principle is: 49.9% & 50.0%
- For Naive Bayes the training and testing accuracies using third party libraries are: 69.8% & 69% and using the first principle is: 69.8% & 69%.

Conclusion:

Draw conclusions from your experiments

- Feature Selection:
  - Feature selection with kbest and PCA doesn't give any clear recommendation about which model is better as both of the models gives almost same result when comparing logistic and SVM accuracies at: 86% and 95% respectively. But while comparing Naive Bayes kbest method gives higher accuracy at 69% compared to 60% when using PCA.
- Model Selection:
  - Logistic regression provided maximum accuracy for training and testing when using both first principle python languages and third party lbibraries at 74.8% & 74.4% with 86% & 84% respectively.
- First Principle vs Third party libraries:
  - Naive bayes model from scratch provided most similar result then other models, so we can conclude that naive bayes from scratch was the best when comes to first principle language, although in most of the cases third party languages provided much higher accuracies.
- Cross Validation:
  - Using cross validation did not provided much improvement in the model, as all the accuracies remains almost similar.
- SVM:
  - The support vector machine model provides the least accuracy at 50% compared to its alternative version using sklearn, which gives accuracy at 95%.
- Further Improvements:
  - To further improve the moddel we could have used hyper parater tuning or different machine learning techniques, that might have improved our results.