News Classification Using Natural Language Processing

Abstract:

News is a thing which the user wants to know and it frequently recommends depending on user preferences. If news is structured in social network, then this classification analyse which group is spreading news. Each news is categorized and updated frequently. For every news there is a main heading of topic, based on this the news is categorized. The media would know which news users are interested. The type of news is changed based on situation.

Objective:

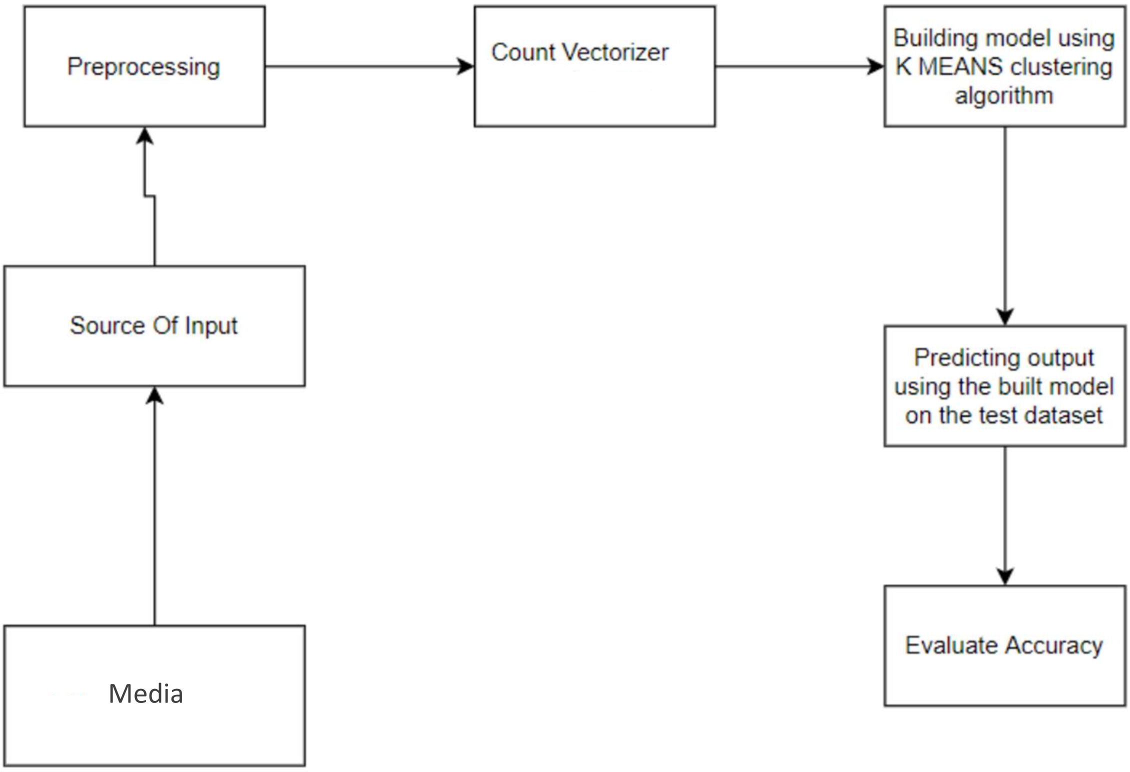
The main objective of this project is to identify the news which was spreading in social media is true or false. In social media, there are many news were spreading and users will believe the rumoured news also. So, this News Classification will give whether the news is real or fake. Because of this the users cannot face any fake on the topic.

Introduction:

Many News Media look at good platform to advertise their news on social media. Many news were spreading on media from different organizations in our daily routine, but many times it becomes a hectic to decide which one is true and which one is fake. Every news that we read is not true. The classifications of news were occasionally revised since the reporter and the reader had different viewpoints. The News Classification system helps to seek out the realism of the news. If the news isn't real, then the user is suggested with the applicable article.

Methodology:

The proposed system when subjected to a scenario of a set of news articles, the new articles are categorized as true or fake by the existing data available. This prediction is done by using the relationship between the words used in the article with one another. The proposed system contains a Vectorization model for finding the relationship between the words and with the obtained information of the existing relations, the new articles are categorized into fake and real news.


Preprocessing
Count Vectorizer
Building model using K MEANS clustering algorithm
Source Of Input
Predicting output using the built model on the test dataset
Evaluate Accuracy
Media

Input is collected from various sources such as newspapers, social media and stored in datasets. System will take input from datasets. The datasets undergo pre-processing and the unnecessary information is removed from it and the data types of the columns are changed if required. Count vectorizer technique is used in the initial step. For fake news detection, we have to train the system using dataset.

Before entering to the detection of fake news, entire dataset is divided into two datasets. 80% is used for training and 20% is used for testing. During training, K-Means algorithm is used to train the model using the train dataset. In testing, the test dataset is given as input and the output is predicted. After the testing time, the predicted output and the actual output are compared using confusion matrix obtained. The confusion matrix gives the information regarding the number of correct and wrong predictions in the case of real and fake news. The accuracy is calculated by the equation No of Correct Predictions/Total Test Dataset Input Size.

CODE:

- The code was done one python jupyter notebook.
- All the coding images are below:

In [1]:
```
pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\srava\anaconda3\lib\site-packages
(3.7)
Requirement already satisfied: click in c:\users\srava\anaconda3\lib\site-packages
(from nltk) (8.0.4)
Requirement already satisfied: regex>=2021.8.3 in c:\users\srava\anaconda3\lib\sit
e-packages (from nltk) (2022.7.9)
Requirement already satisfied: joblib in c:\users\srava\anaconda3\lib\site-package
s (from nltk) (1.1.0)
Requirement already satisfied: tqdm in c:\users\srava\anaconda3\lib\site-packages
(from nltk) (4.64.1)
Requirement already satisfied: colorama in c:\users\srava\anaconda3\lib\site-packa
ges (from click->nltk) (0.4.5)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
# import libraries
import nltk
import pandas as pd
```

In [3]:
```
nltk.download('punkt')
```

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

Out[3]:
```
True
```

import datasets

In [4]:
```
false = pd.read_csv("Fake.csv")
real = pd.read_csv("True.csv")
```

In [5]:
```
false
```

Out[5]:

	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
...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied ‘Safe Zone’ Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016

23481 rows × 4 columns

In [6]:

```
real
```

Out[6]:

	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
...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017

21417 rows × 4 columns

In [7]:

```
false["fact"]=0
real["fact"]=1
```

In [8]:

```
data = pd.concat([false,real], axis=0)
```

In [9]:

```
data
```

Out[9]:

	title	text	subject	date	fact
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

44898 rows × 5 columns

In [10]:

```
data=data.reset_index(drop=True)
data=data.drop(["title","subject","date"],axis=1)
```

Data Preprocessing

In [11]:

```
#Tokenization
from nltk.tokenize import word_tokenize
data['text']= data['text'].apply(word_tokenize)
```

In [12]:

```
#stemming
from nltk.stem.snowball import SnowballStemmer
sbs= SnowballStemmer('english',ignore_stopwords=False)
```

In [13]:

```
def stem_it(text):
    return [sbs.stem(word) for word in text]
```

In [14]:

```
data['text']= data['text'].apply(stem_it)
```

In [15]:

```
def stopword_remover(text):
    return [word for word in text if len(word)>>2]
```

In [16]:

```
data['text']= data['text'].apply(' '.join)
data
```

Out[16]:

	text	fact
0	donald trump just couldn t wish all american a...	0
1	hous intellig committe chairman devin nune is ...	0
2	on friday , it was reveal that former milwauke...	0
3	on christma day , donald trump announc that he...	0
4	pope franci use his annual christma day messag...	0
...	...	...
44893	brussel (reuter) - nato alli on tuesday welc...	1
44894	london (reuter) - lexisnexi , a provid of le...	1
44895	minsk (reuter) - in the shadow of disus sovi...	1
44896	moscow (reuter) - vatican secretari of state...	1
44897	jakarta (reuter) - indonesia will buy 11 suk...	1

44898 rows × 2 columns

Splitting data set

In [17]:

```
from sklearn.model_selection import train_test_split
```

In [18]:

```
X_train, X_test, y_train, y_test = train_test_split(data['text'],
                                                     data['fact'], test_size=0.25)
```

In [19]:

```
X_train
```

Out[19]:

```
18800    the set for this episod is yale univers , a ca...
25181    los angel ( reuter ) - the emmi award show was...
28475    washington ( reuter ) - u.s. repres steve king...
18585    a georgia middl school scienc teacher and a pa...
36107    pari ( reuter ) - french presid emmanuel macro...
                               ...                        
33736    ( reuter ) - demonstr briefli shut down an ari...
14631    so far , the suprem court has not prevent one ...
40427    brussel ( reuter ) - european commiss presid j...
5494     this is whi we love the notori rbg.suprem cour...
29679    washington ( reuter ) - short after donald tru...
Name: text, Length: 33673, dtype: object
```

Vectorization (TFIDF)

In [20]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf= TfidfVectorizer(max_df=0.7)
tfidf_train=tfidf.fit_transform(X_train)
tfidf_test=tfidf.transform(X_test)
```

In [21]:

```
from sklearn.linear_model import LogisticRegression
mdl1=LogisticRegression(max_iter=900)
mdl1.fit(tfidf_train,y_train)
```

Out[21]:
```
LogisticRegression(max_iter=900)
```

In [22]:
```
y_test
```

Out[22]:
```
24766    1
3040     0
39170    1
20166    0
5475     0
        ..
33513    1
11815    0
890      0
7332     0
36172    1
Name: fact, Length: 11225, dtype: int64
```

In [23]:
```
pred1=mdl1.predict(tfidf_test)
pred1
```

Out[23]:
```
array([1, 0, 1, ..., 0, 0, 1], dtype=int64)
```

In [24]:
```
from sklearn.metrics import confusion_matrix, accuracy_score
sc1=accuracy_score(y_test,pred1)
sc1
```

Out[24]:
```
0.987706013363029
```

In [25]:
```
from sklearn.metrics import classification_report
print("Confusion Matrix: ",confusion_matrix(y_test, pred1))
print ("Accuracy : ",accuracy_score(y_test,pred1)*100)
print("Report : ",classification_report(y_test,pred1))
```

```
Confusion Matrix:  [[5792   84]
 [  54 5295]]
Accuracy :  98.7706013363029
Report :                precision    recall  f1-score   support

           0       0.99      0.99      0.99      5876
           1       0.98      0.99      0.99      5349

    accuracy                           0.99     11225
   macro avg       0.99      0.99      0.99     11225
weighted avg       0.99      0.99      0.99     11225
```

In [26]:
```
from sklearn.linear_model import PassiveAggressiveClassifier
mdl2=PassiveAggressiveClassifier(max_iter=100)
mdl2.fit(tfidf_train,y_train)
```

Out[26]:
```
PassiveAggressiveClassifier(max_iter=100)
```

In [27]:
```
pred2=mdl2.predict(tfidf_test)
pred2
```

Out[27]:
```
array([1, 0, 1, ..., 0, 0, 1], dtype=int64)
```

In [28]:
```
sc2=accuracy_score(y_test,pred2)
sc2
```

Out[28]:
```
0.9956347438752784
```

Conclusion:

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. Hence, we used passive aggressive and TF-IDF Vectorizer which is efficient and effective way to obtain accurate results.

The goal of this project is to comprehensively review, summarize, compare and evaluate the current research on fake news. After applying the above algorithms, we can easily classify if the given user input article is real or fake. Fake news infested environment, as the concept of fraud detection in social media is still relatively new. We can help people make more informed decisions this way, and they won't be tricked into believing what others want them to believe.