Tittle - News Classification using Natural Language Processing

ABSTRACT:
News Classification using Natural Language Processing (NLP) is an important application of machine learning. The goal of this project is to classify news articles as either genuine or fake based on their text content. We utilized various NLP techniques and machine learning algorithms to achieve high accuracy in classification. The project was implemented using Python and popular libraries such as Pandas, NLTK, and Scikit-learn.

OBJECTIVE:
The objective of this project is to develop a system that can effectively classify news articles as genuine or fake. The system will use NLP techniques to analyse the text content of the articles and then apply machine learning algorithms to make accurate classifications. The project aims to achieve high accuracy in classification by leveraging the power of modern machine learning techniques.

INTRODUCTION:

With the widespread availability of the internet, news articles can be easily disseminated to a global audience. However, not all news articles are genuine, and many are intentionally misleading or completely fabricated. This has led to an increased demand for systems that can accurately classify news articles as genuine or fake. Machine learning techniques have shown great promise in achieving this goal by analysing the text content of news articles and making accurate classifications.

METHODOLOGY:
The dataset used for this project was obtained from curated collections of genuine and fake news articles. After pre-processing the data, we applied NLP techniques to analyse the text content of the articles. This involved tokenization and stemming of the words in the text content, and removal of stop words. We used the Snowball Stemmer algorithm from the NLTK library to perform stemming.
We then used the TfidfVectorizer algorithm from Scikit-learn to convert the text content of the articles into feature vectors that can be used as input to machine learning algorithms. We split the dataset into training and testing sets and applied two machine learning algorithms: Logistic Regression and Passive Aggressive Classifier. We trained these algorithms using the training set and evaluated their performance on the testing set.

CODE:

```
import nltk
import pandas as pd
fake = pd.read_csv(r"E:\ASUSPC\Downloads\Fake News Data Set.csv",
low_memory=False)


genuine = pd.read_csv(r"E:\ASUSPC\Downloads\True News Data Set.csv")
```

```
print("fake news data shape: ", fake.shape)
print("genuine news data shape: ", genuine.shape)

fake["genuineness"] = 0
genuine["genuineness"] = 1
data = pd.concat([fake, genuine], axis=0)
data = data.reset_index(drop=True)
data = data.drop(['title', 'subject', 'date'], axis = 1)

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

from nltk.stem.snowball import SnowballStemmer
sb = SnowballStemmer("english", ignore_stopwords=False)

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

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

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

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

print("pre processed text: ")
print(data.head(15))




from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =
train_test_split(data['text'],data['genuineness'], test_size = 0.25)

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)

from sklearn.linear_model import LogisticRegression
model1 = LogisticRegression(max_iter=900)
model1.fit(tfidf_train, y_train)

pred1 = model1.predict(tfidf_test)

from sklearn.metrics import accuracy_score
cr1 = accuracy_score(y_test, pred1)
print("Logistic Regression model accuracy:", cr1)

from sklearn.linear_model import PassiveAggressiveClassifier
model2 = PassiveAggressiveClassifier(max_iter=100)
model2.fit(tfidf_train, y_train)

pred2 = model2.predict(tfidf_test)
cr2 = accuracy_score(y_test, pred2)
print("Passive Aggressive Classifier model accuracy:", cr2)
```

CONCLUSION:

In this project, we developed a system that can accurately classify news articles as genuine or fake using NLP techniques and machine learning algorithms. Our system achieved high accuracy in classification, demonstrating the effectiveness of the approach. The system can be used to automatically classify news articles and identify potentially misleading or fabricated content.

OUTPUT:

```
C:\Users\risha\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/User
fake news data shape:  (23502, 172)
genuine news data shape:  (21417, 4)
pre processed text:
                                                text  ... genuineness
0   donaldtrumpjustcouldntwishallamericanahappinew...  ...           0
1   housintelligcommittechairmandevinnuneisgotohav...  ...           0
2   onfriday,itwasrevealthatformermilwaukesheriffd...  ...           0
3   onchristmaday,donaldtrumpannouncthathewouldbeb...  ...           0
4   popefranciusehisannualchristmadaymessagtorebuk...  ...           0
5   thenumberofcaseofcopbrutalandkillpeoplofcolors...  ...           0
6   donaldtrumpspentagoodportionofhisdayathisgolfc...  ...           0
7   inthewakeofyetanothcourtdecisthatderaildonaldt...  ...           0
8   manipeoplhaveraisthealarmregardthefactthatdona...  ...           0
9   justwhenyoumighthavethoughtwedgetabreakfromwat...  ...           0
10  acenterpiecofdonaldtrumpscampaign,andnowhispre...  ...           0
11  republicanareworkovertimtritosellheirscamofat...  ...           0
12  republicanhavehadsevenyeartocomeupwithaviablre...  ...           0
13  themediahasbeentalkalldayabouttrumpandtherepub...  ...           0
14  abigaildisneyisanheiresswithbrassovariwhowillp...  ...           0

[15 rows x 170 columns]
Logistic Regression model accuracy: 0.9949243098842386
Passive Aggressive Classifier model accuracy: 0.9960819234194123

Process finished with exit code 0
```

```
C:\Users\risha\PycharmProjects\pythonProject\venv\Scripts\python.exe C:/Users/risha/AppData
fake news data shape:  (23502, 172)
genuine news data shape:  (21417, 4)
pre processed text:
                                                    text  ... genuineness
0      donaldtrumpjustcouldntwishallamericanahappinew...  ...           0
1      housintelligcommittechairmandevinnuneisgotohav...  ...           0
2      onfriday,itwasrevealthatformermilwaukesheriffd...  ...           0
3      onchristmaday,donaldtrumpannouncthathewouldbeb...  ...           0
4      popefranciusehisannualchristmadaymessagtorebuk...  ...           0
...                                                  ...  ...         ...
44899  berlin(reuter)-theleaderofgermanissocialdemocr...  ...           1
44900  shanghai(reuter)-anoldreviewofanacademmonograp...  ...           1
44901  dubai(reuter)-a14-year-oldboywhowasdetainbysau...  ...           1
44902  london(reuter)-abduldaoudspiltmostofthecappucc...  ...           1
44903  buenoair(reuter)-argentinasmainlaboruniontookt...  ...           1

[44904 rows x 170 columns]
Logistic Regression model accuracy: 0.9959928762243989
Passive Aggressive Classifier model accuracy: 0.9965271593944791

Process finished with exit code 0
```