Artificial Intelligence Project

Title:

News Classification using Natural Language Processing

Abstract:

Natural Language Processing techniques are used in this project to categorize news stories. The code examines the articles' content, collects pertinent information, and then applies a classification algorithm to assign a category. The outcome is a precise news classification system that can be applied to media monitoring and customized news recommendations.

Objective:

The objective is to make use of Natural Language Processing tools to classify news stories according to their content for use in practical applications and to train and assess the effectiveness of the classification model, using a sizable dataset.

Introduction:

Being up to date on the most recent news has become challenging for people in the modern digital age due to the abundance of fake news content available online. Systems for news recommendations have been developed to offer readers personalized news based on their interests, but before doing so, these systems must reliably categorize news articles into different groups. Natural language processing (NLP) enters the picture in this situation. It is feasible to examine the text of news stories using NLP techniques and extract pertinent attributes to correctly categorise them.

This project aims to develop a news classification system using Natural Language Processing (NLP) techniques to accurately categorize news articles based on their content. A large dataset of news articles will be collected and used to train and evaluate the performance of the classification model. The outcome will be an accurate and scalable news classification system that can be used for personalized news recommendation systems and media monitoring tools.

Methodology:

- **NLP Libraries:** For this project, NLTK (Natural Processing Tool Kit) is the library used which is an essential tool for NLP and ML. It supports tasks such as classification, stemming, tagging, semantic reasoning, and tokenization in Python.
- **Data Collection:** A large dataset of news articles from various sources and categories is collected in excel sheet format.
- **Text Pre-processing:** The collected news articles are pre-processed to remove unwanted elements such as stop words, punctuation, and special characters. This helps improve the accuracy of the classification model.
- **Tokenization:** The pre-processed news articles will be tokenized, i.e. large piece of continuous text is divided into distinct units or tokens so that they can easily be processed by the NLP algorithms.
- **Stemming:** This helps remove suffixes of tokens and reduce different forms of a word to a core root.
- **Stopword removal:** This helps remove the most commonly used words that the search engine has been programmed to ignore.

- **Vectorization:** Using this technique textual data or here, the tokenized data is converted into a numerical format. A matrix is created where each column represents a feature and each row represents an individual review.
- **Classification Algorithms:**
    1. Logistic Regression: This is an ML algorithm used for the classification of problems and is a predictive analysis algorithm based on the concept of probability. The logistic function normalizes everything between 0 and 1 and then interprets the received number as a probability.
    2. Passive Aggressive: It is one of the available incremental learning algorithms and very simple to implement. Passive: If the prediction is correct, keep the model and do not make any changes. Aggressive: If the prediction is incorrect, make changes to the model.
- **Model evaluation:** The performance of the classification model will be evaluated using various metrics such as accuracy, precision, recall, and F1-score. The evaluation will be carried out using a test set of news articles that were not used for training the model.

Code:

#Importing Libraries

```
pip install nltk

import nltk

nltk.download()

import pandas as pd
```

#Data Collection

```
fake = pd.read_csv("Fake.csv")

genuine = pd.read_csv("True.csv")

display(fake.info())

display(genuine.info())

display(genuine.head(10))

display(fake.subject.value_counts())

display(genuine.subject.value_counts())

fake['target'] = 0

genuine['target'] = 1

display(genuine.head(10))

display(fake.head(10))

data = pd.concat([fake,genuine],axis=0)

date=data.reset_index(drop=True)
```

```python
data=data.drop(['subject','date','title'],axis=1)

print(data.columns)
```

#Tokenization

```python
from nltk.tokenize import word_tokenize

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

print(data.head(10))
```

#Stemming

```python
from nltk.stem.snowball import SnowballStemmer

porter = SnowballStemmer("english")

def stem_it(text):

    return [porter.stem(word) for word in text]

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

print(data.head(10))
```

#Stopword removal

```python
def stop_it(t):

    dt=[word for word in t if len(word)>2]

    return dt

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

print(data.head(10))

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

from sklearn.model_selection import train_test_split

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

display(X_train.head())

print('\n')

display(y_train.head())
```

#Vectorization

```python
from sklearn.feature_extraction.text import TfidfVectorizer

my_tfidf = TfidfVectorizer(max_df=0.7)

tfidf_train = my_tfidf.fit_transform(X_train)

tfidf_test = my_tfidf.transform(X_test)

print(tfidf_train)
```

#Logistic Regression algorithm

```
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score

model_1=LogisticRegression(max_iter=900)

model_1.fit(tfidf_train,y_train)

pred_1 = model_1.predict(tfidf_test)

cr1    = accuracy_score(y_test,pred_1)

print(cr1*100)
```

#Passive Aggressive algorithm

```
from sklearn.linear_model import PassiveAggressiveClassifier

model = PassiveAggressiveClassifier(max_iter=50)
```

#Model Evaluation

```
model.fit(tfidf_train, y_train)

y_pred = model.predict(tfidf_test)

accscore = accuracy_score(y_test, y_pred)

print('The accuracy of prediction is ',accscore*100)
```

Conclusion:

In conclusion, this project successfully developed a news classification system using Natural Language Processing techniques to accurately categorize news articles based on their content.

In addition to emphasizing the value of gathering and processing huge datasets for creating precise machine-learning models, this study demonstrates the effectiveness of NLP techniques in addressing practical issues. By offering consumers accurate and pertinent news information based on their interests and preferences, the proposed news classification system has the potential to completely change the way people consume news.

Screenshots:

Jupyter news-classification Last Checkpoint: 01/18/2023 (unsaved changes)   Logout

File   Edit   View   Insert   Cell   Kernel   Widgets   Help   Not Trusted   Python 3 (ipykernel) O

```
In [1]: pip install nltk

Requirement already satisfied: nltk in c:\users\ayush\anaconda3\lib\site-packages (3.6.5)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: click in c:\users\ayush\anaconda3\lib\site-packages (from nltk) (8.0.3)
Requirement already satisfied: joblib in c:\users\ayush\anaconda3\lib\site-packages (from nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\ayush\anaconda3\lib\site-packages (from nltk) (2021.8.3)
Requirement already satisfied: tqdm in c:\users\ayush\anaconda3\lib\site-packages (from nltk) (4.62.3)
Requirement already satisfied: colorama in c:\users\ayush\anaconda3\lib\site-packages (from click->nltk) (0.4.4)
```

```
In [2]: import nltk
```

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

showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

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

```
In [4]: import pandas as pd
```

```
In [5]: fake = pd.read_csv("Fake.csv")
        genuine = pd.read_csv("True.csv")
```

```
In [7]: display(fake.info())
```

Jupyter news-classification Last Checkpoint: 01/18/2023 (autosaved)   Logout

File   Edit   View   Insert   Cell   Kernel   Widgets   Help   Not Trusted   Python 3 (ipykernel) O

```
In [6]: display(fake.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    23481 non-null  object
 1   text     23481 non-null  object
 2   subject  23481 non-null  object
 3   date     23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB

None
```

```
In [7]: display(genuine.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    21417 non-null  object
 1   text     21417 non-null  object
 2   subject  21417 non-null  object
 3   date     21417 non-null  object
dtypes: object(4)
memory usage: 669.4+ KB

None
```

Jupyter news-classification Last Checkpoint: 01/18/2023 (unsaved changes)   Logout

File   Edit   View   Insert   Cell   Kernel   Widgets   Help   Not Trusted   Python 3 (ipykernel) O

```
In [8]: display(genuine.head(10))
```

	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
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017
6	Trump says Russia probe will be fair, but time...	WEST PALM BEACH, Fla (Reuters) - President Don...	politicsNews	December 29, 2017
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	The following statements were posted to the ve...	politicsNews	December 29, 2017
8	Trump on Twitter (Dec 28) - Global Warming	The following statements were posted to the ve...	politicsNews	December 29, 2017
9	Alabama official to certify Senator-elect Jone...	WASHINGTON (Reuters) - Alabama Secretary of St...	politicsNews	December 28, 2017

```
In [9]: display(fake.subject.value_counts())

News               9050
politics           6841
left-news          4459
Government News     1570
US_News              783
Middle-east          778
Name: subject, dtype: int64
```

```
In [8]: display(genuine.head(10))
```

	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
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017
6	Trump says Russia probe will be fair, but time...	WEST PALM BEACH, Fla (Reuters) - President Don...	politicsNews	December 29, 2017
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	The following statements were posted to the ve...	politicsNews	December 29, 2017
8	Trump on Twitter (Dec 28) - Global Warming	The following statements were posted to the ve...	politicsNews	December 29, 2017
9	Alabama official to certify Senator-elect Jone...	WASHINGTON (Reuters) - Alabama Secretary of St...	politicsNews	December 28, 2017

```
In [9]: display(fake.subject.value_counts())

        News               9050
        politics           6841
        left-news          4459
        Government News     1570
        US_News             783
        Middle-east         778
        Name: subject, dtype: int64
```

```
In [10]: display(genuine.subject.value_counts())

         politicsNews    11272
         worldnews       10145
         Name: subject, dtype: int64
```

```
In [11]: fake['target'] = 0
         genuine['target'] = 1
```

```
In [12]: display(genuine.head(10))
```

	title	text	subject	date	target
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
5	White House, Congress prepare for talks on spe...	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - T...	politicsNews	December 29, 2017	1
6	Trump says Russia probe will be fair, but time...	WEST PALM BEACH, Fla (Reuters) - President Don...	politicsNews	December 29, 2017	1
7	Factbox: Trump on Twitter (Dec 29) - Approval ...	The following statements were posted to the ve...	politicsNews	December 29, 2017	1
8	Trump on Twitter (Dec 28) - Global Warming	The following statements were posted to the ve...	politicsNews	December 29, 2017	1
9	Alabama official to certify Senator-elect Jone...	WASHINGTON (Reuters) - Alabama Secretary of St...	politicsNews	December 28, 2017	1

```
In [13]: display(fake.head(10))
```

	title	text	subject	date	target
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
5	Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017	0
6	Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017	0
7	Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017	0
8	Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017	0
9	WATCH: Brand New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017	0

```
In [14]: data = pd.concat([fake,genuine],axis=0)
```

```
In [15]: date=data.reset_index(drop=True)
```

```
In [16]: data=data.drop(['subject','date','title'],axis=1)
```

```
In [17]: print(data.columns)

         Index(['text', 'target'], dtype='object')
```

```python
In [17]: print(data.columns)
```
```
Index(['text', 'target'], dtype='object')
```

```python
In [18]: from nltk.tokenize import word_tokenize
```

```python
In [19]: data['text']=data['text'].apply(word_tokenize)
```

```python
In [20]: print(data.head(10))
```
```
                                                text  target
0  [Donald, Trump, just, couldn, t, wish, all, Am...       0
1  [House, Intelligence, Committee, Chairman, Dev...       0
2  [On, Friday, ,, it, was, revealed, that, forme...       0
3  [On, Christmas, day, ,, Donald, Trump, announc...       0
4  [Pope, Francis, used, his, annual, Christmas, ...       0
5  [The, number, of, cases, of, cops, brutalizing...       0
6  [Donald, Trump, spent, a, good, portion, of, h...       0
7  [In, the, wake, of, yet, another, court, decis...       0
8  [Many, people, have, raised, the, alarm, regar...       0
9  [Just, when, you, might, have, thought, we, d,...       0
```

```python
In [21]: from nltk.stem.snowball import SnowballStemmer
         porter = SnowballStemmer("english")
```

```python
In [22]: def stem_it(text):
             return [porter.stem(word) for word in text]
```

```
9  [Just, when, you, might, have, thought, we, d,...       0
```

```python
In [21]: from nltk.stem.snowball import SnowballStemmer
         porter = SnowballStemmer("english")
```

```python
In [22]: def stem_it(text):
             return [porter.stem(word) for word in text]
```

```python
In [23]: data['text']=data['text'].apply(stem_it)
```

```python
In [24]: print(data.head(10))
```
```
                                                text  target
0  [donald, trump, just, couldn, t, wish, all, am...       0
1  [hous, intellig, committe, chairman, devin, nu...       0
2  [on, friday, ,, it, was, reveal, that, former,...       0
3  [on, christma, day, ,, donald, trump, announc,...       0
4  [pope, franci, use, his, annual, christma, day...       0
5  [the, number, of, case, of, cop, brutal, and, ...       0
6  [donald, trump, spent, a, good, portion, of, h...       0
7  [in, the, wake, of, yet, anoth, court, decis, ...       0
8  [mani, peopl, have, rais, the, alarm, regard, ...       0
9  [just, when, you, might, have, thought, we, d,...       0
```

```python
In [25]: def stop_it(t):
             dt=[word for word in t if len(word)>2]
             return dt
```

```python
In [26]: data['text']=data['text'].apply(stop_it)
```

```python
In [26]: data['text']=data['text'].apply(stop_it)
```

```python
In [27]: print(data.head(10))
```
```
                                                text  target
0  [donald, trump, just, couldn, wish, all, ameri...       0
1  [hous, intellig, committe, chairman, devin, nu...       0
2  [friday, was, reveal, that, former, milwauke, ...       0
3  [christma, day, donald, trump, announc, that, ...       0
4  [pope, franci, use, his, annual, christma, day...       0
5  [the, number, case, cop, brutal, and, kill, pe...       0
6  [donald, trump, spent, good, portion, his, day...       0
7  [the, wake, yet, anoth, court, decis, that, de...       0
8  [mani, peopl, have, rais, the, alarm, regard, ...       0
9  [just, when, you, might, have, thought, get, b...       0
```

```python
In [28]: data['text']=data['text'].apply(' '.join)
```

```python
In [29]: from sklearn.model_selection import train_test_split
         X_train,X_test,y_train,y_test=train_test_split(data['text'],data['target'], test_size=0.25)
         display(X_train.head())
         print('\n')
         display(y_train.head())
```
```
22890    21st centuri wire say .from may day riot desig...
19208    can anyon imagin sit republican senat call for...
8209     consid left-lean msnbc still sign joe scarboro...
10151    washington reuter presid barack obama monday 1...
10355    jess watter gop can drain the swamp when their...
Name: text, dtype: object
```

Jupyter news-classification Last Checkpoint: 01/18/2023 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted | Python 3 (ipykernel) ○

```python
In [29]: from sklearn.model_selection import train_test_split
         X_train,X_test,y_train,y_test=train_test_split(data['text'],data['target'], test_size=0.25)
         display(X_train.head())
         print('\n')
         display(y_train.head())
```

```
22890    21st centuri wire say .from may day riot desig...
19208    can anyon imagin sit republican senat call for...
8209     consid left-lean msnbc still sign joe scarboro...
10151    washington reuter presid barack obama monday 1...
10355    jess watter gop can drain the swamp when their...
Name: text, dtype: object


22890    0
19208    0
8209     0
10151    1
10355    0
Name: target, dtype: int64
```

```python
In [30]: from sklearn.feature_extraction.text import TfidfVectorizer
         my_tfidf = TfidfVectorizer(max_df=0.7)

         tfidf_train = my_tfidf.fit_transform(X_train)
         tfidf_test = my_tfidf.transform(X_test)
```

```python
In [31]: print(tfidf_train)
```

```
(0, 83842)    0.10506180073366947
(0, 53670)    0.05676994271158182
(0, 13526)    0.06370224814239053
(0, 78620)    0.10642313351004637
(0, 89630)    0.04985603097045665
(0, 61305)    0.04696157008632665
(0, 31875)    0.11468968857035121
(0, 55618)    0.03814401609104735
(0, 68148)    0.06909424701536585
(0, 65987)    0.10325568545882793
(0, 40372)    0.15687686688803044
(0, 82157)    0.07001813135386722
(0, 16198)    0.14097906339580635
(0, 49807)    0.09079895749957058
(0, 35279)    0.24982391536178666
(0, 51076)    0.11382192986261715
(0, 47085)    0.08921164919385904
(0, 59139)    0.07238806018191683
(0, 79571)    0.08293612540641093
(0, 32188)    0.09569469103465986
(0, 30995)    0.09317050005238989
(0, 19540)    0.12059967505877188
(0, 53387)    0.11657341021218971
(0, 51454)    0.09654328200891465
(0, 62965)    0.14166007291939217
  :    :
(33672, 13199)    0.033578863683986826
(33672, 81431)    0.02315749514610429 2
(33672, 49582)    0.03209640847614171
```

```
(33672, 72493)    0.014544665519547086 4
```

```python
In [32]: from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import accuracy_score
```

```python
In [33]: model_1=LogisticRegression(max_iter=900)
         model_1.fit(tfidf_train,y_train)
         pred_1 = model_1.predict(tfidf_test)
         cr1    = accuracy_score(y_test,pred_1)
         print(cr1*100)
```

```
98.78841870824053
```

```python
In [34]: from sklearn.linear_model import PassiveAggressiveClassifier

         model = PassiveAggressiveClassifier(max_iter=50)
         model.fit(tfidf_train, y_train)
```

```
Out[34]: PassiveAggressiveClassifier(max_iter=50)
```

```python
In [35]: y_pred = model.predict(tfidf_test)
         accscore = accuracy_score(y_test, y_pred)
         print('The accuracy of prediction is ',accscore*100)
```

```
The accuracy of prediction is  99.69710467706014
```

```python
In [ ]:
```