FAKE NEWS DETECTION USING NLP

Project: Fake news detection

Author: Athithya.A

**Comprehensive overview:**

To identify bogus news, sentiment analysis using NLP can be an effective strategy. NLP algorithms can ascertain the intention and any biases of an author by analyzing the emotions displayed in a news story or social media post. Fake news frequently preys on readers' emotions by using strong language or exaggeration.

**Numpy and Pandas for Data Cleaning**. It could have been better to perform initial data cleaning /data processing more faster if I used spreadsheet.But the only reason why I used python libraries for data cleaning or data processing is that I can clearly demonstrate each and every step of data cleaning use it.For an instance,how the data has been converted from its original format into the correct format.

**Geophy for feature Engineering(**created a new field using information from already existing fields which can have scopes of usage in future.)

**Matplotlib,**seaborn and follum for general exploratory data analysis and for some basic data visualization within the jupyter note book.

**Tableau** for visualizations of key finding /Insights and for creating the final dashboard

NOTE: This notebooks includes only data cleaning ,data transformation(creating a new data frame using records from scopeful rows),Feature Engineering ,few EDA and some visualizations.

**Problem definition:**

The problem is to develop a fake news detection model using a Kaggle dataset. The goal is to distinguish between genuine and fake news articles based on their titles and text. This project involves using natural language processing (NLP) techniques to preprocess the text data, building a machine learning model for clas sification, and evaluating the model's performance.Fake news detection is the process of identifying whether a given piece of information, typically in the form of news articles or claims, is false or misleading. With the rise of social media and user-generated content, the dissemination of false or misleading information has become widespread, leading to a pressing need for automated tools to combat this challenge. Natural Language Processing (NLP) offers powerful techniques to understand and analyze textual data to determine its authenticity.

Design Thinking:

**Datasource:** Detecting fake news using Natural Language Processing (NLP) requires a combination of high-quality datasets, robust algorithms, and domain expertise. When you're looking to build or train a model to detect fake news, starting with a good dataset is crucial. Here are some of the popular datasets and data sources used for fake news detection:

1. **LIAR Dataset**: This dataset contains a detailed annotation of over 12K short statements labeled for their truthfulness, subject, context, etc. The statements are extracted from platforms like Politifact.
2. **FakeNewsNet**: This dataset provides a comprehensive structure of fake news and its spread over social media platforms. It covers news articles and user interactions (tweets/shares) for these news items.
3. **BuzzFeed News Dataset**: BuzzFeed News had manually analyzed and labeled some news stories on their platform during the 2016 U.S. presidential election. These labeled data are available for research purposes.
4. **Kaggle Fake News Dataset**: Kaggle, a platform for data science competitions, has a dataset for fake news detection. This dataset consists of labeled news articles that can be used for binary classification tasks.
5. **Getting Real about Fake News**: This dataset consists of real and fake news stories. The labels for this dataset are based on the evaluations of six independent fact-checkers.
6. **BS Detector**: A dataset derived from the tool that flags articles from questionable sources. It is based on a constantly-updated list of known fake news sources.
7. **FakeNewsTracker**: This is a collection of thousands of articles from various domains and is one of the extensive datasets available for fake news detection.
8. **Fact Checking Websites**: Websites such as Snopes, FactCheck.org, and Politifact occasionally provide datasets or lists of checked claims which can be scraped or accessed in some manner for building a dataset.

When building models for fake news detection:

* **Beware of biases**: The labeling of news as 'fake' or 'real' can be subjective and might reflect certain biases. Always check the source and methodology of the dataset collection to ensure fairness and lack of bias.
* **Combine multiple sources**: To ensure your model is robust and generalizes well, consider amalgamating datasets from multiple sources or curating your own dataset by pooling data from various sources.
* **Domain expertise**: Fake news detection isn't just about spotting linguistic patterns. Understanding the topic, context, and source is equally crucial. Combining NLP techniques with domain expertise can often yield better results.
* **Continual learning**: The style and strategy of fake news propagation can evolve. It's essential to keep updating and training your models to adapt to the latest trends.
* *# This Python 3 environment comes with many helpful analytics libraries installed*
* *# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python*
* *# For example, here's several helpful packages to load*
* import warnings
* warnings.filterwarnings('ignore')
* import numpy as np *# linear algebra*
* import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)*
* import nltk
* from nltk.sentiment import SentimentIntensityAnalyzer
* import warnings
* warnings.filterwarnings("ignore")
* from sklearn.model\_selection import train\_test\_split
* from sklearn.feature\_extraction.text import TfidfVectorizer
* from sklearn.svm import SVC
* from sklearn.metrics import accuracy\_score, classification\_report
* *# Input data files are available in the read-only "../input/" directory*
* *# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory*
* import os
* for dirname, \_, filenames **in** os.walk('/kaggle/input'):
* for filename **in** filenames:
* print(os.path.join(dirname, filename))
* *# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"*
* *# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session*
* /kaggle/input/fake-and-real-news-dataset/True.csv
* /kaggle/input/fake-and-real-news-dataset/Fake.csv

true = pd.read\_csv('/kaggle/input/fake-and-real-news-dataset/True.csv')

fake = pd.read\_csv('/kaggle/input/fake-and-real-news-dataset/Fake.csv')

In [3]:

fake['Category'] = 'fake'

fake

Out[3]:

|  | title | text | subject | date | Category |
| --- | --- | --- | --- | --- | --- |
| 0 | Donald Trump Sends Out Embarrassing New Year’... | Donald Trump just couldn t wish all Americans ... | News | December 31, 2017 | fake |
| 1 | Drunk Bragging Trump Staffer Started Russian ... | House Intelligence Committee Chairman Devin Nu... | News | December 31, 2017 | fake |
| 2 | Sheriff David Clarke Becomes An Internet Joke... | On Friday, it was revealed that former Milwauk... | News | December 30, 2017 | fake |
| 3 | Trump Is So Obsessed He Even Has Obama’s Name... | On Christmas day, Donald Trump announced that ... | News | December 29, 2017 | fake |
| 4 | Pope Francis Just Called Out Donald Trump Dur... | Pope Francis used his annual Christmas Day mes... | News | December 25, 2017 | fake |
| ... | ... | ... | ... | ... | ... |
| 23476 | McPain: John McCain Furious That Iran Treated ... | 21st Century Wire says As 21WIRE reported earl... | Middle-east | January 16, 2016 | fake |
| 23477 | JUSTICE? Yahoo Settles E-mail Privacy Class-ac... | 21st Century Wire says It s a familiar theme. ... | Middle-east | January 16, 2016 | fake |
| 23478 | Sunnistan: US and Allied ‘Safe Zone’ Plan to T... | Patrick Henningsen 21st Century WireRemember ... | Middle-east | January 15, 2016 | fake |
| 23479 | How to Blow $700 Million: Al Jazeera America F... | 21st Century Wire says Al Jazeera America will... | Middle-east | January 14, 2016 | fake |
| 23480 | 10 U.S. Navy Sailors Held by Iranian Military ... | 21st Century Wire says As 21WIRE predicted in ... | Middle-east | January 12, 2016 | fake |

23481 rows × 5 columns

In [4]:

true['Category'] = 'true'

true

Out[4]:

|  | title | text | subject | date | Category |
| --- | --- | --- | --- | --- | --- |
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | politicsNews | December 31, 2017 | true |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | politicsNews | December 29, 2017 | true |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | politicsNews | December 31, 2017 | true |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | politicsNews | December 30, 2017 | true |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | politicsNews | December 29, 2017 | true |
| ... | ... | ... | ... | ... | ... |
| 21412 | 'Fully committed' NATO backs new U.S. approach... | BRUSSELS (Reuters) - NATO allies on Tuesday we... | worldnews | August 22, 2017 | true |
| 21413 | LexisNexis withdrew two products from Chinese ... | LONDON (Reuters) - LexisNexis, a provider of l... | worldnews | August 22, 2017 | true |
| 21414 | Minsk cultural hub becomes haven from authorities | MINSK (Reuters) - In the shadow of disused Sov... | worldnews | August 22, 2017 | true |
| 21415 | Vatican upbeat on possibility of Pope Francis ... | MOSCOW (Reuters) - Vatican Secretary of State ... | worldnews | August 22, 2017 | true |
| 21416 | Indonesia to buy $1.14 billion worth of Russia... | JAKARTA (Reuters) - Indonesia will buy 11 Sukh... | worldnews | August 22, 2017 | true |

**Data preprocessing:**

Detecting fake news using NLP requires a robust data preprocessing pipeline since the quality of the input data has a significant impact on the performance of the model. Here's a step-by-step guide for data preprocessing in the context of fake news detection:

1. **Collection of Data**:
   * Gather a substantial amount of news articles labeled as 'real' or 'fake'.
   * Ensure diversity in the data to account for different styles, topics, and sources of news.
2. **Text Cleaning**:
   * **Remove Noise**: Get rid of any URLs, HTML tags, non-textual content, and unnecessary white spaces.
   * **Lowercasing**: Convert the entire text to lowercase to maintain uniformity and reduce the size of the vocabulary.
   * **Tokenization**: Convert sentences into individual words or tokens.
   * **Stop Word Removal**: Remove common words such as 'and', 'the', 'is', etc. which don't add significant meaning in this context.
   * **Stemming/Lemmatization**: Reduce words to their root form. For instance, 'running' becomes 'run'. Lemmatization is a more sophisticated approach than stemming, as it looks into the morphology of the words.
3. **Handling Imbalanced Data**:
   * Fake news detection datasets might be imbalanced with more real news articles than fake ones or vice versa.
   * Strategies:
   * **Undersampling**: Remove some samples from the overrepresented class **Oversampling**: Add more samples to the underrepresented class, for instance using the SMOTE algorithm.

CODE:

data['Category'].value\_counts()

Out[7]:

Category

fake 23481

true 21417

Name: count, dtype: int64

In [8]:

*#Transforming category values to numerical*

from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

data['Category'] = encoder.fit\_transform(data['Category'])

In [9]:

data['Category']

Out[9]:

0 0

1 0

2 0

3 0

4 0

..

44893 1

44894 1

44895 1

44896 1

44897 1

Name: Category, Length: 44898, dtype: int64

In [10]:

vectorizer = TfidfVectorizer()

title = vectorizer.fit\_transform(data['title'])

title

Out[10]:

<44898x20896 sparse matrix of type '<class 'numpy.float64'>'

with 546512 stored elements in Compressed Sparse Row format>

Feature extraction:

1. **Bag of Words (BoW)**: This is the simplest method where you create a vocabulary of unique words from the entire corpus of texts and represent each article as a vector indicating the presence or absence (or frequency) of each word.
2. **TF-IDF (Term Frequency-Inverse Document Frequency)**: Unlike BoW which only considers frequency, TF-IDF also takes into account the importance of a word in a document relative to its frequency across other documents.
3. **Word Embeddings**: Methods like Word2Vec or GloVe create dense vector representations for words based on their semantic meanings. An article can be represented as the average (or other aggregation) of its word vectors.
4. **Sentiment Analysis**: The sentiment of an article might be indicative of its veracity. Fake news might be more polarizing or extreme in sentiment.
5. **Named Entity Recognition (NER)**: Check for entities mentioned in the content (like persons, organizations, and locations) and their consistency or relevance with known facts.
6. **Readability Scores**: Fake news articles might be written differently in terms of complexity or structure. Metrics like the Flesch-Kincaid readability score could be used as features.
7. **Stylometry**: Analyzing the writing style, like sentence length, word length, vocabulary richness, etc. Some fake news creators might have distinctive writing styles.
8. **Topic Modeling**: Techniques like Latent Dirichlet Allocation (LDA) can be used to identify topics in articles. The distribution of topics could be a feature.
9. **N-grams**: Instead of just individual words, consider sequences of 2 (bigrams),

Model selection:

import warnings

warnings.filterwarnings('ignore')

import pandas as pd

import string

import matplotlib.pyplot as plt

import seaborn as sns

import nltk

import plotly.express as px

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score

Model training

In [59]:

df\_Fake = pd.read\_csv(r'/kaggle/input/fake-and-real-news-dataset/Fake.csv')

df\_True = pd.read\_csv(r'/kaggle/input/fake-and-real-news-dataset/True.csv')

In [60]:

df\_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

In [61]:

df\_True.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

In [62]:

df\_Fake.head()

Out[62]:

|  | 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 |

In [63]:

df\_True.head()

Out[63]:

|  | 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 |

In [64]:

df\_Fake['Fake'] = 1

df\_True['Fake'] = 0

In [65]:

df = pd.concat([df\_Fake, df\_True], axis=0, ignore\_index=True)

In [66]:

df['subject'].value\_counts()

Out[66]:

subject

politicsNews 11272

worldnews 10145

News 9050

politics 6841

left-news 4459

Government News 1570

US\_News 783

Middle-east 778

Name: count, dtype: int64

In [67]:

sns.countplot(x='subject', data =df, hue = 'Fake')

plt.xticks(rotation =90)

plt.show()

In [68]:

df['date\_len'] = df['date'].astype(str).apply(lambda x: len(x))

In [69]:

df['date\_len'].value\_counts()

Out[69]:

date\_len

12 10465

17 8099

18 6334

11 4022

16 3927

14 3742

13 3026

15 2730

Evaluation:

It is 99% accurate.