Fake News Detection using Machine Learning

October 8, 2024

1 Fake News Detection using Machine Learning on the WELFake Dataset

1.1 Introduction

In today's digital age, the proliferation of fake news has become a significant concern. The ability to automatically detect and classify news articles as real or fake is crucial for maintaining the integrity of information. This project aims to develop a machine learning model that can accurately classify news articles using the WELFake dataset. This notebook documents the entire process, from data loading and preprocessing to model training, evaluation, and deployment.

1.2 Table of Contents

1. Importing Libraries
2. Loading and Exploring the Data
    - Dataset Description
    - Initial Data Inspection
3. Data Preprocessing
    - Handling Missing Values
    - Exploratory Data Analysis
        - Text Length Analysis
    - Text Cleaning
4. Feature Extraction
    - TF-IDF Vectorization
5. Model Training
    - Train-Test Split
    - Multinomial Naive Bayes Classifier
    - Random Forest Classifier
6. Model Evaluation
    - Classification Reports
7. Model Saving and Deployment
    - Saving Models with Pickle
    - Loading Models for Prediction
8. Prediction on New Data
9. Challenges and Solutions
10. Conclusion

11. Future Work
12. References

1.3 1. Importing Libraries

We begin by importing all the necessary libraries required for data manipulation, visualization, preprocessing, and model building.

[64]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from nltk import word_tokenize
from string import punctuation
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
from pickle import dump
#import nltk
#nltk.download('punkt')
#nltk.download('stopwords')
```

1.4 2. Loading and Exploring the Data

1.4.1 Dataset Description

The WELFake dataset is a comprehensive collection of news articles, merged from four popular datasets (Kaggle, McIntire, Reuters, BuzzFeed Political) to prevent overfitting and provide ample text data for machine learning training.

- **Total Entries:** 72,134 news articles
  - **Real News:** 35,028 articles (Label = 1)
  - **Fake News:** 37,106 articles (Label = 0)
- **Columns:**
  - `Serial number`: Unique identifier for each article
  - `Title`: Headline of the news article
  - `Text`: Main content of the news article
  - `Label`: Indicates whether the news is real (1) or fake (0)

1.4.2 Initial Data Inspection

We load the dataset and perform initial inspections to understand its structure and identify any immediate issues.

[61]:
```
#load Data
data = pd.read_csv("WELFake_Dataset.csv")

print(data.head())
```

```
   Unnamed: 0                                              title  \
0           0  LAW ENFORCEMENT ON HIGH ALERT Following Threat...
1           1                                                NaN
2           2  UNBELIEVABLE! OBAMA’S ATTORNEY GENERAL SAYS MO...
3           3  Bobby Jindal, raised Hindu, uses story of Chri...
4           4  SATAN 2: Russia unvelis an image of its terrif...

                                                text  label
0  No comment is expected from Barack Obama Membe...      1
1     Did they post their votes for Hillary already?      1
2   Now, most of the demonstrators gathered last ...      1
3  A dozen politically active pastors came here f...      0
4  The RS-28 Sarmat missile, dubbed Satan 2, will...      1
```

1.5 3. Data Preprocessing

1.5.1 Handling Missing Values

We check for missing values and handle them appropriately to ensure data integrity.

[62]:
```
data.drop(columns='Unnamed: 0',inplace=True)

print(data.shape)

print(data.isnull().sum())

data.fillna(' ',inplace=True)
print(data.isnull().sum())
```

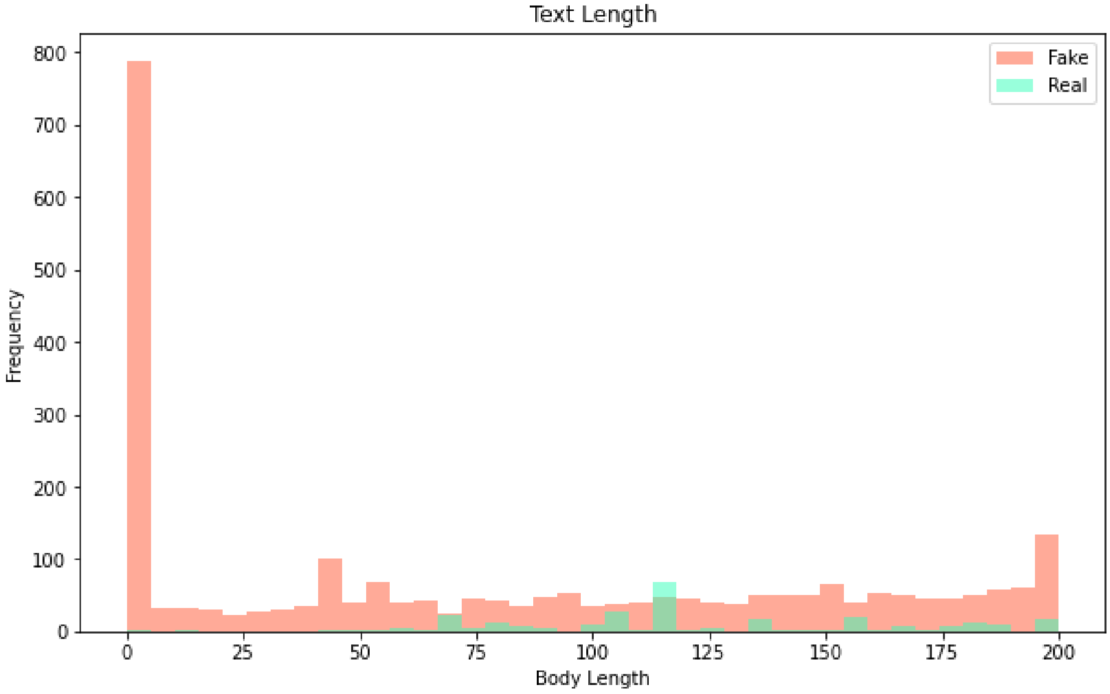
```
(72134, 3)
title    558
text      39
label      0
dtype: int64
title    0
text     0
label    0
dtype: int64
```

1.5.2 Exploratory Data Analysis

Text Length Analysis *We analyze the distribution of text lengths to understand differences between fake and real news articles.*

[65]:
```
data['body_len'] = data['text'].apply(len)

bins = np.linspace(0, 200, 40)
plt.figure(figsize=(10, 6))
plt.hist(data[data["label"]== 1]["body_len"], bins, alpha=0.5, label="Fake",␣
 ↪color="#FF5733")
plt.hist(data[data["label"]== 0]["body_len"], bins, alpha=0.5, label="Real",␣
 ↪color="#33FFB8")
plt.legend(loc="upper left")
plt.xlabel('Body Length')
plt.ylabel('Frequency')
plt.title('Text Length')
plt.legend(loc='upper right')
plt.show()
```



1.5.3 Text Cleaning

We define a function to clean the text data, which includes several preprocessing steps to prepare the data for vectorization.

Tokenization, Stopword Removal, and Stemming

[ ]:
```
ps = PorterStemmer()
def clean_text(txt):
        txt = txt.lower()
        txt = word_tokenize(txt)
        txt = [t for t in txt if t not in punctuation]
        txt = [t for t in txt if t not in stopwords.words("english")]
        txt = [ps.stem(t)for t in txt]
        txt = " ".join(txt)
        return txt


data.loc[:,"clean_text"]=data["text"].apply(clean_text)
print(data.head())
```

1.6 4. Feature Extraction

1.6.1 TF-IDF Vectorization

We transform the cleaned text data into numerical features using TF-IDF vectorization, which considers both term frequency and inverse document frequency.

[16]:
```
cv = TfidfVectorizer()
vector = cv.fit_transform(c_data["clean_text"])
```

[17]:
```
print(vector.shape)
```

```
(71537, 211485)
```

[18]:
```
features = pd.DataFrame.sparse.from_spmatrix(vector,columns=cv.
 ↪get_feature_names_out())
target = data['label']
```

1.7 5. Model Training

1.7.1 Train-Test Split

We split the dataset into training and testing sets to evaluate the model's performance on unseen data.

[20]:
```
x_train,x_test,y_train,y_test = train_test_split(features,target)
```

1.7.2 Multinomial Naive Bayes Classifier

We train a Multinomial Naive Bayes classifier, which is suitable for text classification tasks.

[21]:
```
mnb=MultinomialNB()
mnb.fit(x_train,y_train)
```

[21]:
```
MultinomialNB()
```

1.7.3 Random Forest Classifier

We train a Random Forest classifier with 300 estimators to improve prediction accuracy.

[44]:

```
rf = RandomForestClassifier(n_estimators=300)
rf.fit(x_train,y_train)
```

[44]:

```
RandomForestClassifier(n_estimators=300)
```

1.8 6. Model Evaluation

1.8.1 Classification Reports

We evaluate both models using classification reports to compare their performance.

[45]:

```
crnb = classification_report(y_test,mnb.predict(x_test))
crf = classification_report(y_test,rf.predict(x_test))
print(crnb,crf)
```

```
              precision    recall  f1-score   support

           0       0.90      0.92      0.91      8807
           1       0.92      0.90      0.91      9078

    accuracy                           0.91     17885
   macro avg       0.91      0.91      0.91     17885
weighted avg       0.91      0.91      0.91     17885
               precision    recall  f1-score   support

           0       0.94      0.94      0.94      8807
           1       0.94      0.95      0.94      9078

    accuracy                           0.94     17885
   macro avg       0.94      0.94      0.94     17885
weighted avg       0.94      0.94      0.94     17885
```

1.9 7. Model Saving and Deployment

1.9.1 Saving Models with Pickle

We save the trained models and vectorizer using the pickle module for future use without retraining.

[36]:

```
# vector creation
f= open("CV_FRN.pkl","wb")
dump(cv,f)
f.close()

# MultinomialNB model creation
f= open("MNB_FRN.pkl","wb")
dump(mnb,f)
f.close()
```

```
# RandomForestClassifier model creation
f= open("RF_FRN.pkl","wb")
dump(rf,f)
f.close()
```

1.9.2 Loading Models for Prediction

We demonstrate how to load the saved models and vectorizer to make predictions on new data.

[38]:

```
# laod in vector file for prediction
from pickle import load

f=open("CV_FRN.pkl","rb")
cv=load(f)
f.close()

f=open("MNB_FRN.pkl","rb")
mnb=load(f)
f.close()

f=open("RF_FRN.pkl","rb")
rf=load(f)
f.close()
```

1.10 8. Prediction on New Data

We accept user input, preprocess it, and use both models to predict whether the news is fake or real.

[58]:

```
# Prediction

news = input("enter news text ")

# Cleaned user input data
cnews=clean_text(news)

# vectorize cleaned data
vnews=cv.transform([cnews])

# predict using both Model
#MultinomialNB model
pred_mnb=mnb.predict(vnews)

#RandomForestClassifier model
pred_rf=rf.predict(vnews)

print(pred_rf[0],pred_mnb[0])
```

enter news text A dozen politically active pastors came here for a private dinner Friday night to hear a conversion story unique in the context of presidential politics: how Louisiana Gov. Bobby Jindal traveled from Hinduism to Protestant Christianity and, ultimately, became what he calls an “evangelical Catholic.” Over two hours, Jindal, 42, recalled talking with a girl in high school who wanted to “save my soul,” reading the Bible in a closet so his parents would not see him and feeling a stir while watching a movie during his senior year that depicted Jesus on the cross. “I was struck, and struck hard,” Jindal told the pastors. “This was the Son of God, and He had died for our sins.” Jindal’s session with the Christian clergy, who lead congregations in the early presidential battleground states of Iowa and South Carolina, was part of a behind-the-scenes effort by the Louisiana governor to find a political base that could help propel him into the top tier of Republican candidates seeking to run for the White House in 2016. Known in GOP circles mostly for his mastery of policy issues such as health care, Jindal, a Rhodes Scholar and graduate of the Ivy League’s Brown University, does not have an obvious pool of activist supporters to help drive excitement outside his home state. So he is harnessing his religious experience in a way that has begun to appeal to parts of the GOP’s influential core of religious conservatives, many of whom have yet to find a favorite among the Republicans eyeing the presidential race. Other potential 2016 GOP candidates are wooing the evangelical base, including Sens. Rand Paul (Ky.) and Ted Cruz (Tex.) and Indiana Gov. Mike Pence. But over the weekend in Lynchburg - a mecca of sorts for evangelicals as the home of Liberty University, founded in the 1970s by the Rev. Jerry Falwell - Jindal appeared to make progress. In addition to his dinner with the pastors, he delivered a well-received “call to action” address to 40,000 Christian conservatives gathered for Liberty’s commencement ceremony, talking again about his faith while assailing what he said was President Obama’s record of attacking religious liberty. The pastors who came to meet Jindal said his intimate descriptions of his experiences stood out. “He has the convictions, and he has what it takes to communicate them,” said Brad Sherman of Solid Rock Christian Church in Coralville, Iowa. Sherman helped former Arkansas governor Mike Huckabee in his winning 2008 campaign for delegates in Iowa. Another Huckabee admirer, the Rev. C. Mitchell Brooks of Second Baptist Church in Belton, S.C., said Jindal’s commitment to Christian values and his compelling story put him “on a par” with Huckabee, who was a Baptist preacher before entering politics. The visiting pastors flew to Lynchburg over the weekend at the invitation of the American Renewal Project, a well-funded nonprofit group that encourages evangelical Christians to engage in the civic arena with voter guides, get-out-the-vote drives and programs to train pastors in grass-roots activism. The group’s founder, David Lane, has built a pastor network in politically important states such as Iowa, Missouri, Ohio and South Carolina and has led trips to Israel with Paul and others seeking to make inroads with evangelical activists. The group that Lane invited to Lynchburg included Donald Wildmon, a retired minister and founder of the American Family Association, a prominent evangelical activist group that has influence through its network of more than 140 Christian radio stations. Most of the pastors that Lane’s organization brought to Lynchburg had not met Jindal. But they said he captured their interest recently when he stepped forward to defend Phil Robertson, patriarch of the ‘‘Duck Dynasty’’ television-show family, amid a controversy over disparaging remarks he made about gays in an interview with GQ magazine. Throughout his Lynchburg visit, Jindal presented himself as a willing culture warrior. During his commencement address Saturday, he took up the cause of twin brothers whose HGTV reality series about renovating and reselling houses, ‘‘Flip It Forward,’’ was canceled last week after a Web site revealed that they had protested against same-sex marriage at the 2012 Democratic National Convention in Charlotte. The siblings, Jason and David Benham, both Liberty graduates, attended the graduation and a private lunch with Jindal, who called the action against them ‘‘another demonstration of intolerance from the entertainment industry.’’ ‘‘If these guys had protested at the Republican Party convention, instead of canceling their show, HGTV would probably have given them a raise,’’ Jindal said as the Liberty crowd applauded. He cited the Hobby Lobby craft store chain, which faced a legal challenge after refusing to provide employees with insurance coverage for contraceptives as required under the Affordable Care Act. Members of the family that owns Hobby Lobby, who have become heroes to many religious conservatives, have said that they are morally opposed to the use of certain types of birth control and that they considered the requirement a violation of their First Amendment right to religious freedom. The family was ‘‘committed to honor the Lord by being generous employers, paying well above minimum wage and increasing salaries four years in a row even in the midst of the enduring recession,’’ Jindal told the Liberty graduates. ‘‘None of this matters to the Obama administration.’’ But for the pastors who came to see Jindal in action, the governor’s own story was the highlight of the weekend. And in many ways, he was unlike any other aspiring president these activists had met. Piyush Jindal was born in 1971, four months after his parents arrived in Baton Rouge, La., from their native India. He changed his name to Bobby as a young boy, adopting the name of a character on a favorite television show, ‘‘The Brady Bunch.’’ His decision to become a Christian, he told the pastors, did not come in one moment of lightning epiphany. Instead, he said, it happened in phases, growing from small seeds planted over time. Jindal recalled that his closest friend from grade school gave him a Bible with his name emblazoned in gold on the cover as a Christmas present. It struck him initially as an unimpressive gift, Jindal told the pastors. ‘‘Who in the world would spend good money for a Bible when everyone knows you can get one free in any hotel?’’ he recalled thinking at the time. ‘‘And the gold lettering meant I couldn’t give it away or return it.’’ His religious education reached a higher plane during his junior year in high school, he told his dinner audience. He wanted to ask a pretty girl on a date during a hallway conversation, and she started talking about her faith in God and her opposition to abortion. The girl invited him to visit her church. Jindal said he was skeptical and set out to ‘‘investigate all these fanciful claims’’ made by the girl and other friends. He started reading the Bible in his closet at home. ‘‘I was unsure how my parents would react,’’ he said. After the stirring moment when he saw Christ depicted on the cross during the religious movie, the Bible and his very existence suddenly seemed clearer to him, Jindal told the pastors. Jindal did not dwell on his subsequent conversion to Catholicism just a few years later in college, where he said he immersed himself in the traditions of

```
the church. He touched on it briefly during the commencement address, noting in
passing that “I am best described as an evangelical Catholic.” Mostly, he sought
to showcase the ways in which he shares values with other Christian
conservatives. “I read the words of Jesus Christ, and I realized that they were
true,” Jindal told the graduates Saturday, offering a less detailed accounting
of his conversion than he had done the night before with the pastors. “I used to
think that I had found God, but I believe it is more accurate to say that He
found me.”

0 0
```

1.11 9. Challenges and Solutions

During the project, we faced several challenges due to the large size of the dataset and high dimensionality of the feature space.

- **Memory Limitations:**
  - **Issue:** Memory errors occurred when processing n-grams and bigrams with `CountVectorizer`, resulting in over 600,000 features.
  - **Solution:** Switched to `TfidfVectorizer` and used a sparse matrix representation to handle the large feature set efficiently.
- **Processing Time:**
  - **Issue:** Data splitting and model training were time-consuming, with some steps taking several hours.
  - **Solution:** Opted for more efficient algorithms and limited the use of resource-intensive processes.
- **Data Preprocessing Decisions:**
  - **Issue:** Including stopwords led to increased dimensionality and memory issues.
  - **Solution:** Removed stopwords to reduce the feature space and avoid memory constraints.

1.12 10. Conclusion

By carefully preprocessing the data and selecting appropriate models, we successfully built classifiers capable of distinguishing between fake and real news with high accuracy. The Random Forest Classifier performed better, achieving approximately 94% accuracy compared to 91% with Multinomial Naive Bayes. Despite challenges related to memory and processing time, the project demonstrates the effectiveness of machine learning in text classification tasks.

1.13 11. Future Work

To further improve the model and its applicability, we plan to:

- **Optimize Memory Usage:**
  - Explore dimensionality reduction techniques like PCA.
  - Utilize distributed computing methods.
- **Enhance Models:**

    - Experiment with deep learning architectures such as RNNs or Transformers.
    - Implement cross-validation for model robustness.
- **Advanced Feature Engineering:**
    - Incorporate word embeddings like Word2Vec or GloVe.
    - Use additional metadata features if available.

1.14 12. References

- WELFake Dataset Publication: IEEE Transactions on Computational Social Systems
- NLTK Documentation: NLTK 3.6.2
- Scikit-learn Documentation: scikit-learn
- Python Pickle Module: pickle — Python object serialization