MA208 Probability Theory and Applications

"Fake News Detection using Probabilistic model"

Submitted by:
Chandan Naik(171CO212)
chandannaik999@gmail.com
4th year, Computer Science and Engineering

DEPARTMENT OF MATHEMATICAL AND COMPUTATIONAL SCIENCE
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL
2020-2021

Contents:

Abstract:

Accounting to the expeditious digitization across all channels and mediums, the menace of fake news has been burgeoning at a colossal scale. The majority of the countries all across the world are trying to combat this challenge. This project explores the application of Probability and Machine Learning techniques to identify fake news accurately. Pre-processing tools are used to clean the data and apply feature extraction on them. Then a fake news detection model is built using Naive Bayes to find the best fit for the model.

Problem statement:

This project describes a simple fake news detection method based on one of the artificial intelligence algorithms – naive Bayes classifier. The goal of the project is to examine how this particular method works for this particular problem given a manually labeled news dataset and to support (or not) the idea of using artificial intelligence for fake news detection. The developed system was tested on a relatively new data set, which gave an opportunity to evaluate its performance on recent data.

Motivation:

Over the last decade, there have encounters of flux in misinformation that spread like wildfires. The surge in fake news was noticed during the 2016 presidential elections that happened in the US that determined the fate of these elections. In many cases, it is seen that the sharing of hoax news has been more than that of accurate news. In a massive market like India, the scope of fake news propaganda has been artfully misused by many groups. Researches indicate that Facebook and WhatsApp are the platforms that are utilized for spreading fake news. An approximate one in two Indians has agreed to have received fake news during the 2019 Lok Sabha elections.

It is seen that spam messages and fake news have striking similarities. They use manipulative ways to win over the reader's opinions. Most of them also have grammatical mistakes and they also use a similar restricted set of words among them. Since both the media share such similar properties, we can use similar approaches to detect fake news accurately. One way to tackle fake news is to manually classify news as real or fake. Even though that seems like the simplest solution it is not practical with the jillions of news that get produced to manually label it. Hence, there is a need to look for a pragmatic technical solution to do the same. The proposed method in this research is to exploit the advancement in machine learning. To do the same, the classification model has been trained with a Naive Bayes algorithm to label the data.

Literature review

There are several influential articles about automatic deception detection. Here, I discuss some of the recent and important ones.

In the paper by Shu A et al., they investigated how news can be classified as true or not by focusing on a few attributes that are repeatedly encountered in fake news. In their opinion, these characteristics were based on "psychology and social theories, existing algorithms from a data mining perspective, evaluation metrics, and representative datasets". This paper also analyses the different challenges one will encounter while studying this topic. [1]

The paper is written by Rubin et al. deals with the domain of fake news which is composed of satirical news. Satire news intentionally provides hints revealing its own deception. While fake news wants the readers to believe a false fact, satire news must eventually be understood as a jest. This paper provides an in-depth view of the features of humor and satire news along with the style of the author's reporting. The paper has considered the news articles from twelve contemporary news topics in four different domains which are civics, science, business, and soft news. The paper proposes a Support Vector Machine based algorithm which can detect satire news based on features like Absurdity, Humor, Grammar, and Punctuation. The models achieved an accuracy of 90% and a recall of 84%. The aim is to reduce the negative impact of satire news on readers. [2]

In the paper by Kelly Stahl et al., they have considered past and current techniques for fake news identification in text formats while elucidating how and why news fake exists in any case. This paper incorporates a discussion on how the writing style of a paper can also impact on its classification. They had implemented their project using Naïve Bayes Classifier and Support Vector Machines methods. They had looked into the semantic analysis of the text for classification.[3]

In this paper by Marco L. Delia Vedov et al., they say that "we propose a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing their already high accuracy by up to 4.8%". The proposed model was then tested on a real-time application and they achieved high accuracy by testing it on a FB messenger chatbot. The accuracy achieved by them is close to 82%.[4]

---

[1] "Fake News Detection on Social Media: A Data Mining ...." 7 Aug. 2017, https://arxiv.org/abs/1708.01967. Accessed 31 Oct. 2020.

[2] "Fake News or Truth? Using Satirical Cues to Detect Potentially ...." https://www.aclweb.org/anthology/W16-0802. Accessed 31 Oct. 2020.

[3] "[PDF] Fake news detection in social media | Semantic Scholar." https://www.semanticscholar.org/paper/Fake-news-detection-in-social-media-Stahl/b202b4b7124b774391109dc47a33e17224b12295. Accessed 31 Oct. 2020.

[4] "Automatic Online Fake News Detection Combining Content ...." https://ieeexplore.ieee.org/document/8468301. Accessed 31 Oct. 2020.

Solution

A. Dataset

The corpus of data implemented in this project had around 40000 articles of data.The dataset contains 23000 of fake news dataset and 21000 true new datasets. These articles mainly constituted news about US politics. The dataset obtained on Kaggle was noisy and required cleaning. The main features included in each row of the data were title, text, subject, date, classification of being fake or true. The dataset has the following features:

TITLE
TEXT
SUBJECT
DATE
FAKE/TRUE

Sample dataset:

	title	text	subject	date	True/Fake
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

B. Feature Extraction and Pre-Processing

To start off with the implementation, the data is obtained in raw format which is part of the dataset. This data needs to be pre-processed before we can implement it in the project. The process includes stop-word removal followed by making the entire document in lower case for uniformity. Also, any of the special characters that can cause an anomaly in the document are removed in this process. Stop words are words that are not relevant and have little meaning lexically. These words are most often ignored to not cause any discrepancies to the process of classification. In a sentence like “There is a Bengal tiger.”, the first three words ,’there’, ‘is’ and ,’an’ are stop words and have no significant meaning. These are the words that are usually excluded and the examples are: who, of, a, what, etc.

The ‘article’ which is a combination of the title and the text is then tokenized, i.e; split into tokens. The tokens are then converted to a matrix of token counts using the

CountVectorizer. This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix. This is also called the Bag-of-Words model.

TF-IDF(Term Frequency-Inverse document frequency) is used to solve the problem with highly frequent words dominating in the document (e.g. larger score), but not containing as much "informational content" to the model as rarer but perhaps domain specific words.

C. Model

Naïve Bayes is a conditional probability model which can be used for labeling. The goal is to find a way to predict the class variable (B) using a vector of independent variables (A), i.e., finding the function f: A-->B. In probability terms, the goal is to find P(B|A), i.e., the probability of B belonging to a certain class A. B is generally assumed to be a categorical variable with two or more discrete values (A is commonly known as hypotheses and B is knowns as evidence). It is a mathematically simple way to include contributions of many factors in predicting the class of the next data instance in the testing set. The limitation of
Naive Bayes is that they assume that all features are not dependent on each other. The Naive Bayes rule is based on the theorem formulated by Bayes:

$$p(H \mid E) = \frac{p(E \mid H)\ p(H)}{p(E)}$$

Where,

P(H): The probability of hypothesis H being true. This is known as prior probability.
P(E): The probability of the evidence.
P(E|H): The probability of the evidence given that hypothesis is true.
P(H|E): The probability of the hypothesis given that the evidence is true.

Assumption:

The fundamental Naive Bayes assumption is that each feature makes an:

- Independent
- equal

contribution to the outcome.
Note: The assumptions made by Naive Bayes are not generally correct in real-world situations. In fact, the independence assumption is never correct but often works well in practice.

Naive Bayes Classifier

It is a kind of classifier that works on Bayes theorem. In Bayes classifier prediction of membership, probabilities are made for every class such as the probability of data points

associated with a particular class. The class having maximum probability is appraised as the most suitable class.

This is also referred to as Maximum A Posteriori (MAP).
The MAP for a hypothesis is:

- $MAP\ (H) = \max P((H|E))$
- $MAP\ (H) = \max P((H|E) * (P(H)) / P(E))$
- $MAP\ (H) = \max(P(E|H) * P(H))$

Multinomial Naive Bayes: Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the event model typically used for document classification.
$P\ (E)$ is evidence probability, and it is used to normalize the result. The result will not be affected by removing $P(E)$. In Naive Bayes classifiers, we popularly conclude that all the variables or features are not related to each other. The existence or absence of a variable does not impact the existence or absence of any other variable

Specifically, I use the Multinomial Bayes Classifier.

Multinomial Naive Bayes: In multinomial naive Bayes, feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts.

Programing Code:

Colab link here

In [1]:
```
from google.colab import drive
drive.mount('/content/gdrive', force_remount=True)
```

```
Mounted at /content/gdrive
```

In [2]:
```
!cp /content/gdrive/My\ Drive/True.csv /content/
!cp /content/gdrive/My\ Drive/Fake.csv /content/
!ls
```

```
Fake.csv  gdrive  sample_data  True.csv
```

In [12]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
import os
truenews = pd.read_csv('True.csv') # the true news dataset
fakenews = pd.read_csv('Fake.csv') # fake news dataset
```

In [13]:
```
fakenews.head()
```

Out[13]:

	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 [14]:
```
truenews.head()
```

Out[14]:

	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 [15]:
```
fakenews.describe()
```

Out[15]:

	title	text	subject	date
count	23481	23481	23481	23481
unique	17903	17455	6	1681

	title	text	subject	date
top	MEDIA IGNORES Time That Bill Clinton FIRED His...		News	May 10, 2017
freq	6	626	9050	46

In [16]:
```
truenews.describe()
```

Out[16]:

	title	text	subject	date
count	21417	21417	21417	21417
unique	20826	21192	2	716
top	Factbox: Trump fills top jobs for his administ...	(Reuters) - Highlights for U.S. President Dona...	politicsNews	December 20, 2017
freq	14	8	11272	182

In [17]:
```
truenews['True/Fake']='True'
fakenews['True/Fake']='Fake'
```

In [18]:
```
# Combine the 2 DataFrames into a single data frame
news = pd.concat([truenews, fakenews])
news["Article"] = news["title"] + news["text"]
news.sample(frac = 1) #Shuffle 100%
```

Out[18]:

	title	text	subject	date	True/Fake	Arti
6814	Tennessee Republican Leader Vows Punishing Ta...	It s amazing how quickly free market values ...	News	April 19, 2016	Fake	Tenness Republic Leader Vc Punishing T
12070	KELLYANNE CONWAY On Trump's Terrorism Policy: ...	https://www.youtube.com/watch?v=0cVugq2GbBk	politics	Dec 23, 2016	Fake	KELLYAN CONWAY Trum Terrori Policy
16965	North Korea warns threats a 'big miscalculatio...	SYDNEY (Reuters) - North Korea has sent a lett...	worldnews	October 19, 2017	True	North Ko warns thre a ' miscalculati
3753	Ryan tries to tamp down Comey memo furor, says...	WASHINGTON (Reuters) - U.S. House of Represent...	politicsNews	May 17, 2017	True	Ryan tries tamp do Comey me furor, say
5123	Highlights: The Trump presidency on March 5 at...	(Reuters) - Highlights of the day for U.S. Pre...	politicsNews	March 6, 2017	True	Highligh The Tru presidency March 5 a
...	...	...	...	...	...	
3632	Bernie Sanders Gets Brutally Honest On 'Conan...	Bernie Sanders appeared with Conan O Brien on ...	News	November 30, 2016	Fake	Ber Sanders G Brut Honest 'Cona
1505	House to vote on federal budget next week: Hou...	WASHINGTON (Reuters) - The U.S. House of Repre...	politicsNews	September 28, 2017	True	House to v on fede budget n week: Ho

	title	text	subject	date	True/Fake	Arti
8803	Obama must make new budget request for Iraq tr...	WASHINGTON (Reuters) - The Republican chairman...	politicsNews	July 11, 2016	True	Obama m make n bud request Iraq
18046	LOL! CROOKED and IRRELEVANT HILLARY CLINTON GO...	Hillary Clinton will be back in the spotlight ...	left-news	Aug 28, 2017	Fake	L(CROOK a IRRELEVA HILLA CLINT(G(
13221	BOOM! WATCH VP CANDIDATE TIM KAINE “Skirt” The...	Nice try but it s obvious Tim Kaine is coverin...	politics	Aug 18, 2016	Fake	BOC WATCH CANDIDA TIM KAI “Skirt” Th

44898 rows × 6 columns

In [19]:
```
news.head()
```

Out[19]:

	title	text	subject	date	True/Fake	Article
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	True	As U.S. budget fight looms, Republicans flip t...
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	True	U.S. military to accept transgender recruits o...
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	True	Senior U.S. Republican senator: 'Let Mr. Muell...
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	True	FBI Russia probe helped by Australian diplomat...
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	True	Trump wants Postal Service to charge 'much mor...

In [21]:
```
# Data Cleaning
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
import string
def process_text(s):
    # Check string to see if they are a punctuation
    nopunc = [char for char in s if char not in string.punctuation]

    # Join the characters again to form the string.
    nopunc = ''.join(nopunc)
```

```
    # Convert string to lowercase and remove stopwords
    clean_string = [word for word in nopunc.split() if word.lower() not in st
    return clean_string
```

In [22]:
```
# Tokenize the Article
news['Clean Text'] = news['Article'].apply(process_text)
```

In [23]:
```
news.sample(5)
```

Out[23]:

	title	text	subject	date	True/Fake	Article	Clean Tex
2536	Trump says would be surprised if Iran complian...	WASHINGTON (Reuters) - U.S. President Donald T...	politicsNews	July 26, 2017	True	Trump says would be surprised if Iran complian...	[Trump, says would surprised Iran complian..
1053	U.S. partisan split widening over Russia probe...	WASHINGTON (Reuters) - Top Democrats in the U....	politicsNews	October 24, 2017	True	U.S. partisan split widening over Russia probe...	[US, partisan split, widening Russia probe,..
2412	WATCH: Conservatives Don't Punch Nazis, They ...	If there is one political ideology that every ...	News	February 23, 2017	Fake	WATCH: Conservatives Don't Punch Nazis, They ...	[WATCH Conservatives Don't, Punch Nazis, Gi..
10053	White House finds temporary fix in Zika fundin...	WASHINGTON (Reuters) - The White House said on...	politicsNews	April 6, 2016	True	White House finds temporary fix in Zika fundin...	[White, House finds temporary, fix Zika, fu..
9281	WATCH: MEGHAN MCCAIN RIPS Into Joy Behar For H...	The View co-host, and rabid, liberal, activi...	politics	Dec 4, 2017	Fake	WATCH: MEGHAN MCCAIN RIPS Into Joy Behar For H...	[WATCH MEGHAN MCCAIN RIPS, Joy Behar, Emba..

In [25]:
```
from sklearn.feature_extraction.text import CountVectorizer
bow_transformer = CountVectorizer(analyzer=process_text).fit(news['Clean Text
print(len(bow_transformer.vocabulary_)) #Total vocab words
```

```
39099
```

In [26]:
```
#Bag-of-Words (bow) transform the entire DataFrame of text
news_bow = bow_transformer.transform(news['Clean Text'])
```

In [27]:
```
print('Shape of Sparse Matrix: ', news_bow.shape)
print('Amount of Non-Zero occurences: ', news_bow.nnz)
```

```
Shape of Sparse Matrix:  (44898, 39099)
Amount of Non-Zero occurences:  44898
```

In [28]:
```
sparsity = (100.0 * news_bow.nnz / (news_bow.shape[0] * news_bow.shape[1]))
print('sparsity: {}'.format(round(sparsity)))
```

```
sparsity: 0
```

In [29]:
```
#TF-IDF
from sklearn.feature_extraction.text import TfidfTransformer

tfidf_transformer = TfidfTransformer().fit(news_bow)
news_tfidf = tfidf_transformer.transform(news_bow)
print(news_tfidf.shape)
```

```
(44898, 39099)
```

In [30]:
```
# Train Naive Bayes Model
from sklearn.naive_bayes import MultinomialNB
fakenews_detect_model = MultinomialNB().fit(news_tfidf, news['True/Fake'])
```

In [33]:
```
from sklearn.model_selection import train_test_split

news_train, news_test, text_train, text_test = train_test_split(news['Article

print(len(news_train), len(news_test), len(news_train) + len(news_test))
```

```
31428 13470 44898
```

In [34]:
```
from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ('bow', CountVectorizer(analyzer=process_text)),
    ('tfidf', TfidfTransformer()),
    ('classifier', MultinomialNB()),
])
pipeline.fit(news_train,text_train)
```

Out[34]:
```
Pipeline(memory=None,
         steps=[('bow',
                 CountVectorizer(analyzer=<function process_text at 0x7f2174e
8f378>,
                                 binary=False, decode_error='strict',
                                 dtype=<class 'numpy.int64'>, encoding='utf-
8',
                                 input='content', lowercase=True, max_df=1.0,
                                 max_features=None, min_df=1,
                                 ngram_range=(1, 1), preprocessor=None,
                                 stop_words=None, strip_accents=None,
                                 token_pattern='(?u)\\b\\w\\w+\\b',
                                 tokenizer=None, vocabulary=None)),
                ('tfidf',
                 TfidfTransformer(norm='l2', smooth_idf=True,
                                  sublinear_tf=False, use_idf=True)),
                ('classifier',
                 MultinomialNB(alpha=1.0, class_prior=None, fit_prior=Tru
e))],
         verbose=False)
```

In [35]:
```
prediction = pipeline.predict(news_test)
```

In [36]:
```
print(classification_report(prediction,text_test))
```

```
              precision    recall  f1-score   support

        Fake       0.96      0.98      0.97      6892
        True       0.98      0.95      0.97      6578

    accuracy                           0.97     13470
   macro avg       0.97      0.97      0.97     13470
weighted avg       0.97      0.97      0.97     13470
```

Commands Used

A. Dataset

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
import os
truenews = pd.read_csv('True.csv') # the true news dataset
fakenews = pd.read_csv('Fake.csv') # fake news dataset
```

The above code shows the importing of the True news and Fake news dataset

B. Preprocessing:

Removing stopwords from the articles and tokenize the sentences:

```
# Data Cleaning
def process_text(s):
    nopunc = ''.join([char for char in s if char not in string.punctuation])
    clean_string = [word for word in nopunc.split() if word.lower() not in
stopwords.words('english')]
    return clean_string
news['Clean Text'] = news['Article'].apply(process_text)
```

Change the tokens to a spare matrix of tokens. This is also known as the bag-of-words concept.

```
from sklearn.feature_extraction.text import CountVectorizer
bow_transformer = CountVectorizer(analyzer=process_text).fit(news['Clean
Text'])
news_bow = bow_transformer.transform(news['Clean Text'])
```

To solve the problem of highly frequent words starting to dominate in the document is solved using TF-IFD

In this approach, we rescale the frequency of words by how often they appear in all documents so that the scores for frequent words like “the” that are also frequent across all documents are penalized.

```
from sklearn.feature_extraction.text import TfidfTransformer
```

```
tfidf_transformer = TfidfTransformer().fit(news_bow)
news_tfidf = tfidf_transformer.transform(news_bow)
print(news_tfidf.shape)
```

Fit the dataset to the Multinomial Naive bayes model where `news_tfidf` is the training data set and the `news['True/Fake']` is the ground truth

```
from sklearn.naive_bayes import MultinomialNB
fakenews_detect_model = MultinomialNB().fit(news_tfidf, news['True/Fake'])
```

The above mentioned steps can be minimized by using a pipeline that does multiple tasks one-by-one in a pipeline. This method makes the code concise. We pipeline the CountVectorizer(to convert the tokens to the bag-of-words), then the TfidfTransformer, then the MultinomialNB

```
from sklearn.pipeline import Pipeline
pipeline = Pipeline([
    ('bow', CountVectorizer(analyzer=process_text)),
    ('tfidf', TfidfTransformer()),
    ('classifier', MultinomialNB()),])
pipeline.fit(news_train,text_train)
```

The prediction can be evaluated using the following function call

```
prediction = pipeline.predict(news_test)
```

Results and Discussion:

The following results are obtained:

	precision	recall	f1-score	support
Fake	0.96	0.98	0.97	6892
True	0.98	0.95	0.97	6578
accuracy			0.97	13470
macro avg	0.97	0.97	0.97	13470
weighted avg	0.97	0.97	0.97	13470

Results similar to the state of the art model is obtained with the F1 score of **0.97**

Conclusion

In this project, I have presented a model for fake news detection through the Naive Bayes method.
The model achieves the highest accuracy score of 97%.

Fake news detection is an emerging research area that has a scarce number of datasets. There is no data on real-time news or current affairs. The current model is run against the existing dataset, showing that the model performs well against it.

In future work, news article data can be considered related to recent incidents in the corpus of data. The next step then would be to train the model and analyze how the accuracies vary with the new data to further improve it.

References:

https://www.ijrte.org/wp-content/uploads/papers/v8i1C2/A11660581C219.pdf
https://ieeexplore.ieee.org/document/8100379
https://www.geeksforgeeks.org/naive-bayes-classifiers/