Machine Learning Capstone Project

Machine Learning Engineer Nanodegree

Abdallah El Ghamry

September 2020

Fake News Detector Using Text Classification

1. Definition

1.1 Project Overview

Natural Language Processing (NLP) is one of the most important fields in machine learning. It's extremely useful in real-life applications such as: question-answering, spam-detection, language translator, grammar checkers, and more. One of the most important applications of NLP is fake news detector. Rubin et al. [1] discuss 3 types of fake news:

1. Serious fabrications
2. Large-Scale hoaxes
3. Humorous fake news

The hardest type of fake news to detect is humorous fake news. It can deceive the detector system. It is a real challenge for NLP.

Due to the importance of automatic detecting fake news, several researchers have presented proposed models to solve it like using recurrent neural network (RNN).

We will use fake and real news dataset which is available on *Kaggle* platform [2]. It can be obtained at https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. It's a large dataset that will be suitable to train the model. The article's data is the input.

Each entry in the dataset has the following labels:

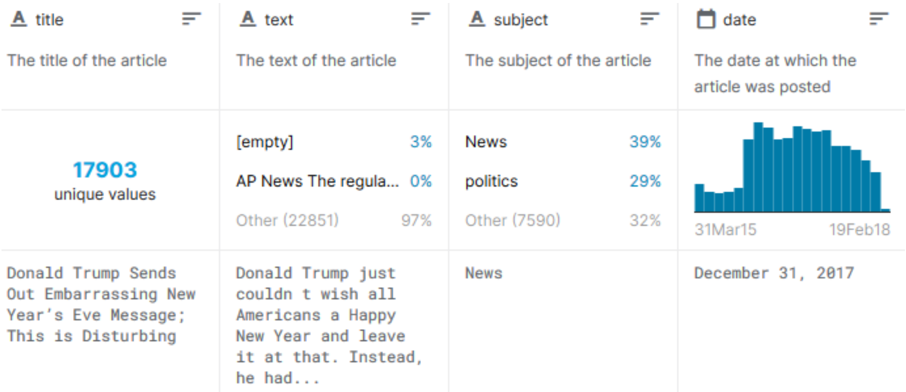
- Title
- Text
- Type (news, politics or others)
- Date (from Mar 2015 to Feb 2018)

- Subject (fake or true)

There are two csv files: *Fake.csv* which represents of the face news, and *True.csv* which represents the real news.

Fake.csv File

- 4 Attributes (title, text, type, date)
- 23481 Records

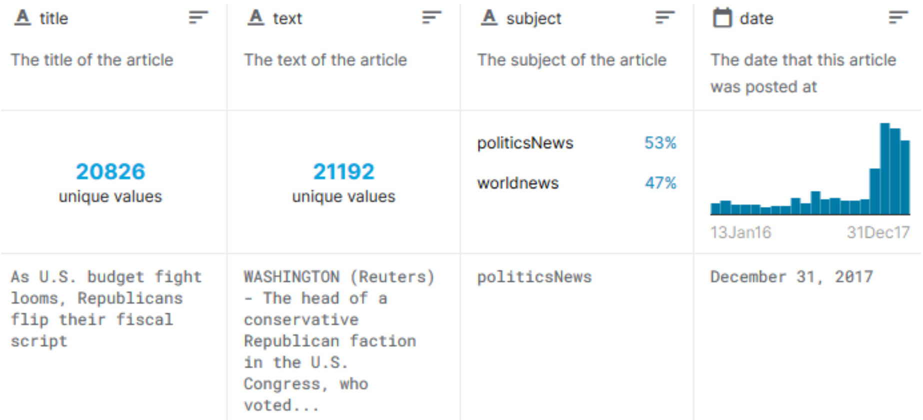


title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date at which the article was posted
17903 unique values	[empty] 3% AP News The regula... 0% Other (22851) 97%	News 39% politics 29% Other (7590) 32%	31Mar15 19Feb18
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had...	News	December 31, 2017

Fig. 1 Fake.csv File Contents

True.csv File

- 4 Attributes (title, text, type, date)
- 21417 Records



title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date that this article was posted at
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	13Jan16 31Dec17
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017

Fig. 2 True.csv File Contents

The personal motivation for selecting this project is fighting rumors. In my country there are many fake news so, I want to develop a model that can detect those fake news.

1.2 Problem Statement

Fake news and articles have become a dangerous effect on online users. Due to the rumors, it is important to detect the fake articles before they spread. The main goal of the model is to find out whether an article is fake or real. This problem is similar to mail filtering and sentiment analysis with some differences.

This problem is a supervised learning. In a supervised learning, the inputs (X) are labeled and the output (y) is labeled. The article is the input and the output is fake or true. A fake article is labeled as 0 while a true article is labeled as 1. Fake News Detection is a binary (0/1) classification problem.

The strategy for solving the problem is described in the following steps:

1. Data exploration and visualization.
2. Data preprocessing.
3. Feature extraction.
4. Splitting the dataset 70% Train set : 30% Test set.
5. Modeling using random forest classifier or any appropriate classifier.
6. Testing the results to evaluate the performance and comparing models.
7. Deploying the model on SageMaker with the best determined accuracy.

An accuracy of at least 95% is the goal. An accuracy of at least 98% is an excellent solution for this problem.

1.3 Metrics

Dividing the correct classifications on the dataset size is a good quantify the performance of a binary classification model. An accuracy of at least 95% was the goal but, a higher accuracy is obtained using our model.

$$accuracy = \frac{correct\ classifications}{cleaned\ dataset\ size} = \frac{true\ positives + true\ negatives}{cleaned\ dataset\ size}$$

A confusion matrix is a table that is used to describe the performance of a classifier.

- True Positives (TP): the true value is 1 and the predicted value is 1.
- True Negatives (TN): the true value is 0 and the predicted value is 0.
- False Positives (FP): the true value is 0 and the predicted value is 1.
- False Negatives (FN): the true value is 1 and the predicted value is 0.

TP	FP
FN	TN

False classifications can appear for several reasons including:

1. Satirical news.
2. Ambiguity of the text or uncleaned text.

2. Analysis

2.1 Data Exploration

As mentioned, the data is stored in two speared files: *True.csv* and *False.csv*. Pandas data frame is a good data structure to represent the dataset. The following table summarizes the basic information of the dataset:

Table 1 Dataset Basic Information

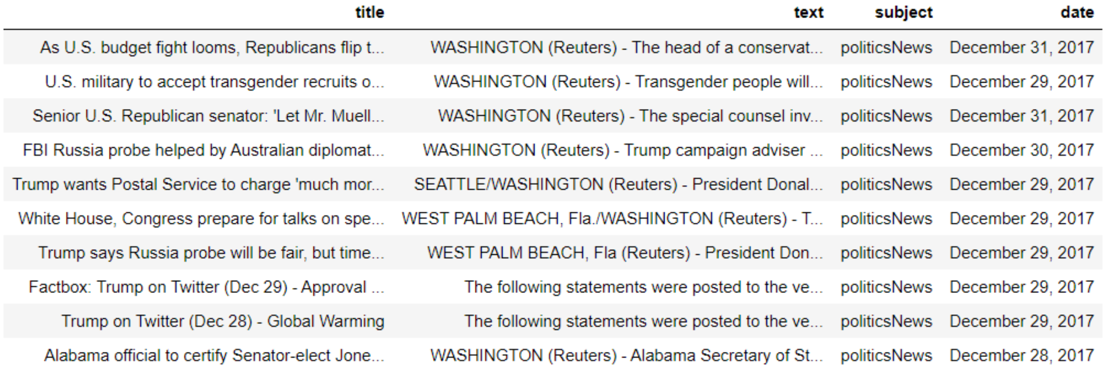
CSV File	Rows	Columns
True.csv	21417	4
False.csv	23481	4

Total Number of Records: 44898

True.csv File:

- There is no null values in the file.
- Types of news: politics news, world news.

Here is a sample of the file:



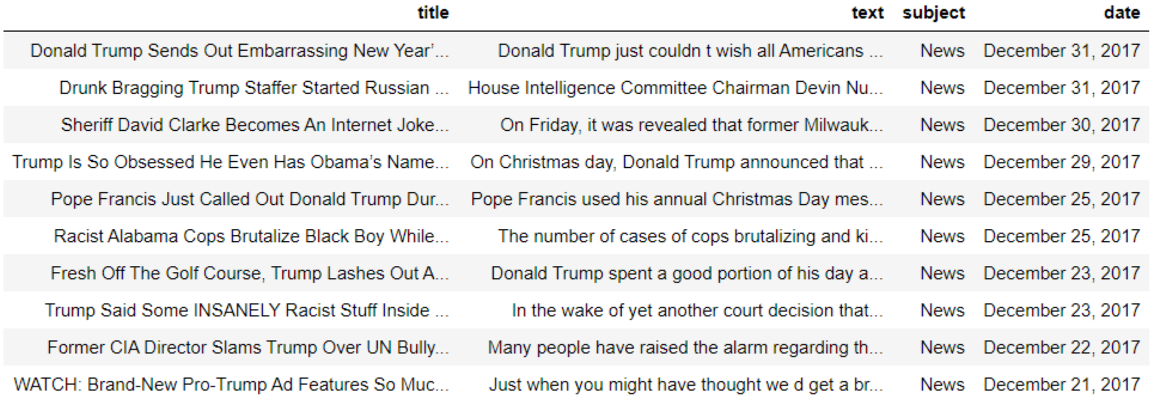
title	text	subject	date
As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
Senior U.S. Republican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

Fig. 3 True.csv Sample

False.csv File:

- There is no null values in the file.
- Types of news: News, politics, Government News, ,left news, US News, Middle east

Here is a sample of the file:



title	text	subject	date
Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017
Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017
Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017
Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017
WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017

Fig. 4 False.csv Sample

Table 2 Dataset Columns

Column	Datatype	Description
Title	String	The title of the article.
Text	String	The body of the article.
Subject	String	The category of the article.
Date	Data	The date of publication.

The most important features are the title and the text of the article. They will be reprocessed in the suitable format.

2.2 Exploratory Visualization

The number of true articles is 21417 while the number of fake articles is 23481.

Fake articles are about 52.3% of the dataset.

True articles are about 47.7% of the dataset.

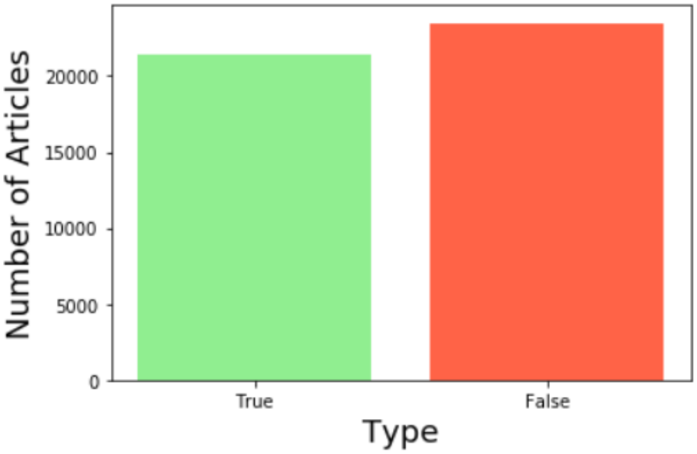


Fig. 5 True Articles vs. False Articles

Political news is the most common news in the dataset. The following figure illustrates the different types of news in the dataset.

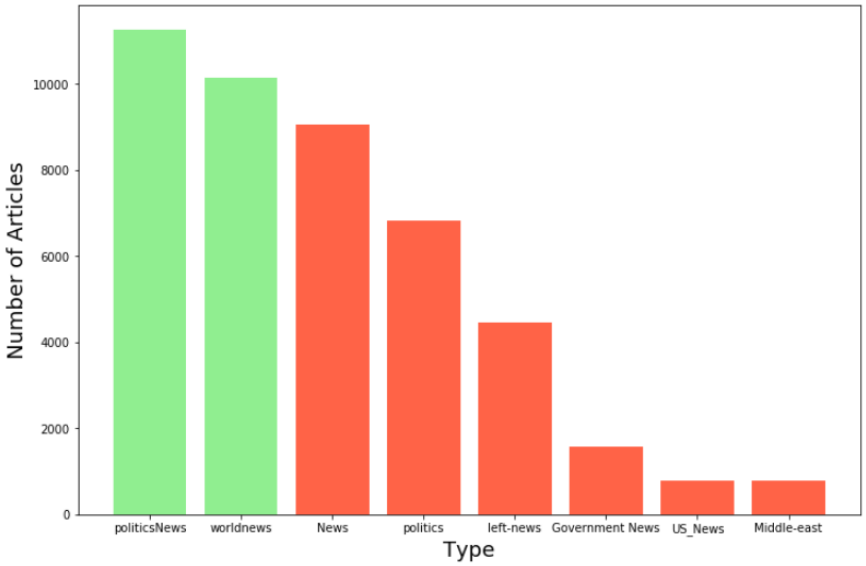


Fig. 6 Article Types

The most common words in the dataset are “trump”, “said”, “state”, “presid”, “u”, “would”, “people”, “year”, “republican”, and “say”. The following figure illustrates the top 10 words in the dataset.

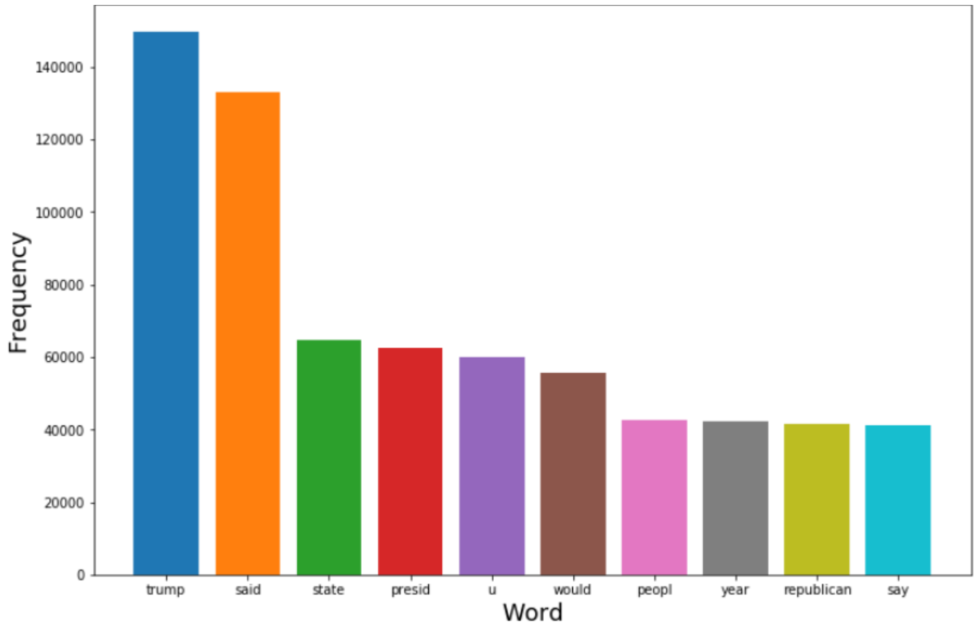


Fig. 7 Top 10 Words in the Dataset

2.3 Algorithms and Techniques

There are many good classifiers such as: Logistic Regression, K-NN, SVM, Naive Bayes, Decision Tree Classification, and Random Forest Classification. A *random forest classifier* from the sklearn.ensemble module is used to fit the data with high accuracy. A random forest classifier is an ensemble learning method for classification. Ensemble is a group of classifiers. Random forest classifier is composed of a number of decision tree classifiers. Random forests perform better one decision tree. Random forest classifier is one of the most powerful Machine Learning algorithms.

Decision Trees they rely on

The parameters of the algorithms are [3]:

- *n_estimators* which is the number of trees in the random forest.
- *criterion* which is the function to measure the quality of a split.
- *random_state* which controls the randomness.

70% of the dataset will be the training set. The final output of the classifier is 0 for fake news and 1 for true news.

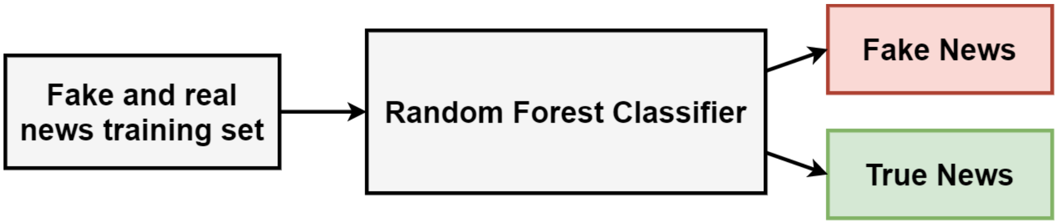


Fig. 8 Training the Random Forest Classifier

The trained classifier will classify the testing set (30% of the dataset) and other news.

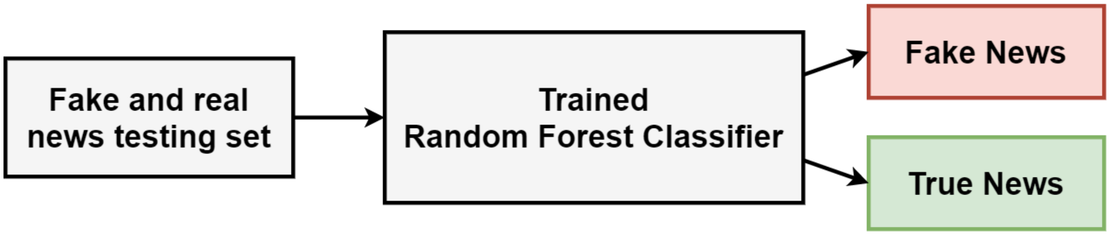

Fake and real news testing set
Trained Random Forest Classifier
Fake News
True News

Fig. 9 Training the Random Forest Classifier

2.4 Benchmark

As mentioned, there are many solutions for this problems on Kaggle platform [2] with high accuracy. So, it will be a challenge to train the model with at least 95% accuracy. The problem will be “Fake News Detection Using RNN” on kaggle [4].

Table 3 Fake News Detection Using RNN Benchmark Results

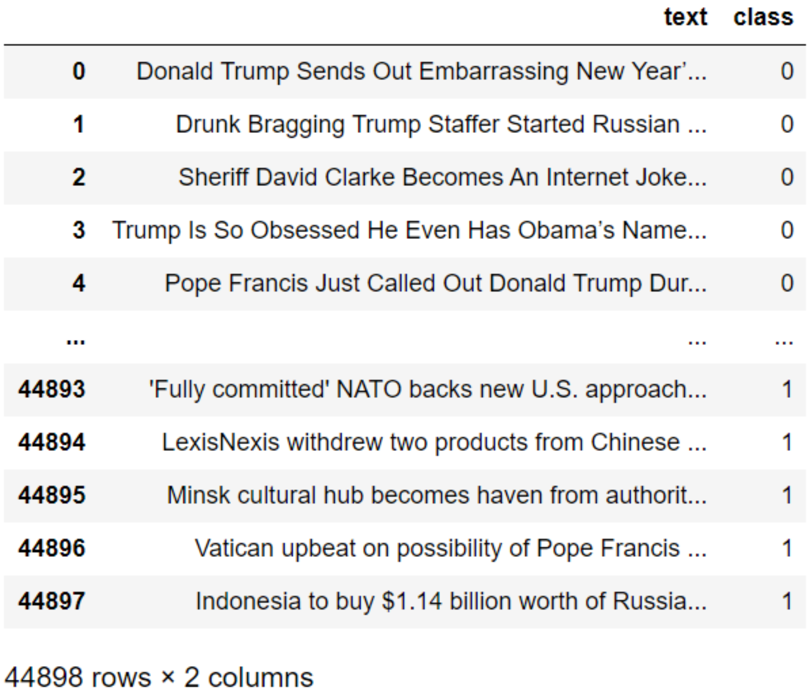
Attribute	Result
Accuracy	0.9903118040089087
Precision	0.9902732746641963
Recall	0.9895857440407313

3. Methodology

3.1 Data Preprocessing

As mentioned, the most important features are the title and the body of the article. The preprocessing steps are:

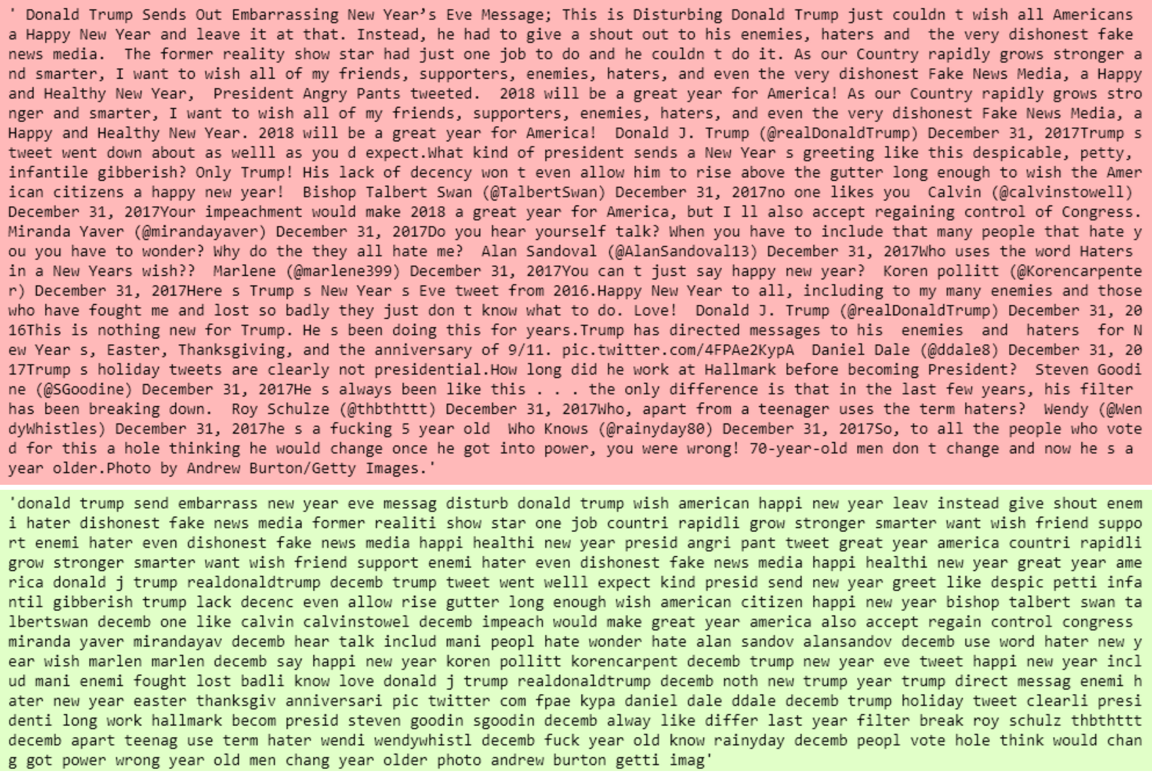
1. Loading the data into two data frames.
   a. df_true represents true news.
   b. df_false represents fake news.
2. Concatenation of the title and the text into one field.
3. Ignoring subject and date fields.
4. Dropping any null values. There is no null values in the dataset but, it is a verification step.
5. Adding an attribute called class (0/1) to classify each entry.
6. Concatenation of the two data frames into a one data frame.
7. Removing unwanted characters. Keeping only alphabet characters (A-Z).
8. Making all of the characters lowercase.
9. Stemming.
10. Removing stop words.



	text	class
0	Donald Trump Sends Out Embarrassing New Year'...	0
1	Drunk Bragging Trump Staffer Started Russian ...	0
2	Sheriff David Clarke Becomes An Internet Joke...	0
3	Trump Is So Obsessed He Even Has Obama's Name...	0
4	Pope Francis Just Called Out Donald Trump Dur...	0
...	...	...
44893	'Fully committed' NATO backs new U.S. approach...	1
44894	LexisNexis withdrew two products from Chinese ...	1
44895	Minsk cultural hub becomes haven from authorit...	1
44896	Vatican upbeat on possibility of Pope Francis ...	1
44897	Indonesia to buy $1.14 billion worth of Russia...	1

44898 rows × 2 columns

Fig. 10 Concatenation of true and false news into one dataset



```
' Donald Trump Sends Out Embarrassing New Year’s Eve Message; This is Disturbing Donald Trump just couldn t wish all Americans
a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and  the very dishonest fake
news media.  The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger a
nd smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy
and Healthy New Year,  President Angry Pants tweeted.  2018 will be a great year for America! As our Country rapidly grows stro
nger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a
Happy and Healthy New Year. 2018 will be a great year for America!  Donald J. Trump (@realDonaldTrump) December 31, 2017Trump s
tweet went down about as welll as you d expect.What kind of president sends a New Year s greeting like this despicable, petty,
infantile gibberish? Only Trump! His lack of decency won t even allow him to rise above the gutter long enough to wish the Amer
ican citizens a happy new year!  Bishop Talbert Swan (@TalbertSwan) December 31, 2017no one likes you  Calvin (@calvinstowell)
December 31, 2017Your impeachment would make 2018 a great year for America, but I ll also accept regaining control of Congress.
Miranda Yaver (@mirandayaver) December 31, 2017Do you hear yourself talk? When you have to include that many people that hate y
ou you have to wonder? Why do the they all hate me?  Alan Sandoval (@AlanSandoval13) December 31, 2017Who uses the word Haters
in a New Years wish??  Marlene (@marlene399) December 31, 2017You can t just say happy new year?  Koren pollitt (@Korencarpente
r) December 31, 2017Here s Trump s New Year s Eve tweet from 2016.Happy New Year to all, including to my many enemies and those
who have fought me and lost so badly they just don t know what to do. Love!  Donald J. Trump (@realDonaldTrump) December 31, 20
16This is nothing new for Trump. He s been doing this for years.Trump has directed messages to his  enemies  and  haters  for N
ew Year s, Easter, Thanksgiving, and the anniversary of 9/11. pic.twitter.com/4FPAe2KypA  Daniel Dale (@ddale8) December 31, 20
17Trump s holiday tweets are clearly not presidential.How long did he work at Hallmark before becoming President?  Steven Goodi
ne (@SGoodine) December 31, 2017He s always been like this . . . the only difference is that in the last few years, his filter
has been breaking down.  Roy Schulze (@thbthttt) December 31, 2017Who, apart from a teenager uses the term haters?  Wendy (@Wen
dyWhistles) December 31, 2017he s a fucking 5 year old  Who Knows (@rainyday80) December 31, 2017So, to all the people who vote
d for this a hole thinking he would change once he got into power, you were wrong! 70-year-old men don t change and now he s a
year older.Photo by Andrew Burton/Getty Images.'
```

```
'donald trump send embarrass new year eve messag disturb donald trump wish american happi new year leav instead give shout enem
i hater dishonest fake news media former realiti show star one job countri rapidli grow stronger smarter want wish friend suppo
rt enemi hater even dishonest fake news media happi healthi new year presid angri pant tweet great year america countri rapidli
grow stronger smarter want wish friend support enemi hater even dishonest fake news media happi healthi new year great year ame
rica donald j trump realdonaldtrump decemb trump tweet went welll expect kind presid send new year greet like despic petti infa
ntil gibberish trump lack decenc even allow rise gutter long enough wish american citizen happi new year bishop talbert swan ta
lbertswan decemb one like calvin calvinstowel decemb impeach would make great year america also accept regain control congress
miranda yaver mirandayav decemb hear talk includ mani peopl hate wonder hate alan sandov alansandov decemb use word hater new y
ear wish marlen marlen decemb say happi new year koren pollitt korencarpent decemb trump new year eve tweet happi new year incl
ud mani enemi fought lost badli know love donald j trump realdonaldtrump decemb noth new trump year trump direct messag enemi h
ater new year easter thanksgiv anniversari pic twitter com fpae kypa daniel dale ddale decemb trump holiday tweet clearli presi
denti long work hallmark becom presid steven goodin sgoodin decemb alway like differ last year filter break roy schulz thbthttt
decemb apart teenag use term hater wendi wendywhistl decemb fuck year old know rainyday decemb peopl vote hole think would chan
g got power wrong year old men chang year older photo andrew burton getti imag'
```

Fig. 11 Data before and after preprocessing

3.2 Implementation

The implementation is in the notebook “FakeNewsDetection.ipynb” and the training file “train/train.pt”

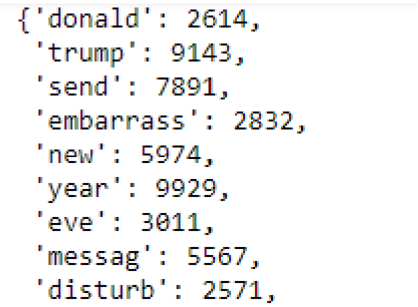
The implementation steps are:

1. Data preprocessing.
2. Feature extraction: each article is represented as 10000 features.
3. Splitting the dataset 70% Train set : 30% Test set.
4. Modeling using random forest classifier with [10 and 20] estimates.
5. Testing the results to evaluate the performance.
6. If the accuracy is less than the wanted accuracy, go to stop 4.

This model is implemented using a random forest classifier. The number of estimators is 20 and the criterion is entropy. Notebook, training, deployment instances are of the type *ml.m4.xlarge*.

```
from sklearn.ensemble import RandomForestClassifier


# Define a model
model =  RandomForestClassifier(n_estimators=20,
                                criterion="entropy",
                                random_state=0)
# Train the model
model.fit(train_x, train_y)
```



```
{'donald': 2614,
 'trump': 9143,
 'send': 7891,
 'embarrass': 2832,
 'new': 5974,
 'year': 9929,
 'eve': 3011,
 'messag': 5567,
 'disturb': 2571,
```

Fig. 12 Vocabulary Sample

The following table summarizes the libraries that are used in the project.

Table 4 Libraries that are used in the project

Library	Usage
numpy	Processing data.
pandas	Loading and reprocessing the dataset.
matplotlib	Visualizing data.
seaborn	Visualizing data.
re	Removing unwanted characters
nltk	Removing stopwords. Stemming.
sklearn	Feature extraction. Splitting the dataset into train and test set. Training the model using random forest classifier. Evaluation of the performance of the model.
sys	Optimizing the main memory.
boto3	Dealing with S3 bucket.
sagemaker	Deploying the model.
os	Accessing data directories.

3.3 Refinement

Increasing the number of estimators could improve the mode. The initial number of estimators was 10. The accuracy was good. It was higher than 98%. The final number of estimators was 20. Twenty estimators achieved a gear accuracy (higher than 99%).

The initial type of the notebook instance was *ml.t2.medium* but, it was not suitable to process this number of data. So, the final instance type was *ml.m4.xlarge*.

Optimizing our resources is so important. Casting the datatype of the features from *int64* to *unit8* will reduce memory utilization. Deleting unused resources also can help to optimize the resources.

Splitting the test set into batches can prevent some errors while the prediction of data because the testing set is huge (13470 article).

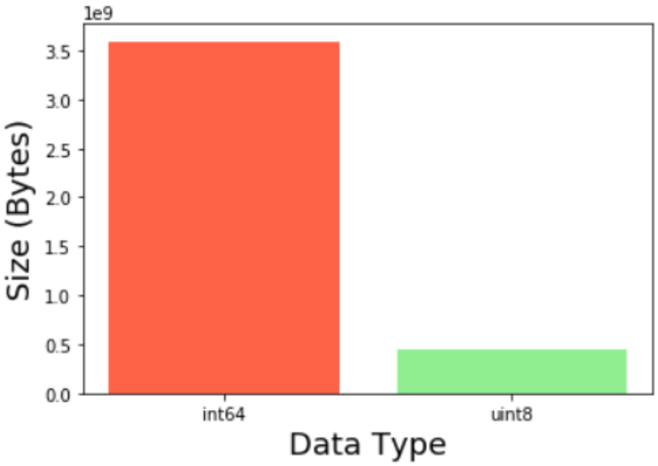


Fig. 13 Memory Optimization

4. Results

4.1 Model Evaluation and Validation

The confusion matrix was chose to evaluate the performance of the model. The final accuracy is 0.9904231625835189 and that is optimal for this problem. Our restrictions states that accuracy must be at least 0.95 while the model accuracy is higher than the expected accuracy. Using the random forest classifier with 20 estimators leads to a great accuracy. Because of the accuracy of the model, we will not try other models to fit the data.

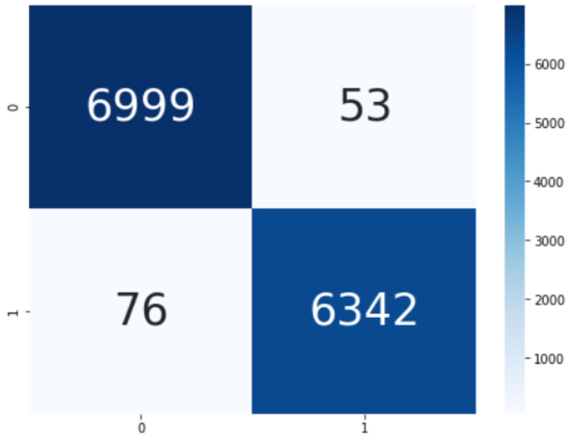


Fig. 13 Confusion Matrix

The following table summarizes the results of the model.

Table 5 Model Results

Attribute	Result
Accuracy	0.9904231625835189
Precision	0.9917122752150117
Recall	0.9881583047678405

4.2 Justification

The results of the benchmark and our models are almost the same. Out model is a little bit better than the benchmark.

Using of the random forest classifier is faster than using recurrent neural network (RRN) model and leads to a little bit better accuracy. The solution significant enough to have adequately solved the problem. The difference between the accuracies is 0.00011135857461019866 which is very small but enough.

Table 6 Models Comparison

Attribute	Benchmark	Our Model	Comment
	RRN	Random Forest	
Accuracy	0.9903118040089087	0.9904231625835189	Higher Accuracy
Precision	0.9902732746641963	0.9917122752150117	Higher Precision
Recall	0.9895857440407313	0.9881583047678405	Lower Recall

References

[1] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: three types of fakes. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST '15). American Society for Information Science, USA, Article 83, 1–4.

[2] Fake and real news dataset. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

[3] Scikit-learn random forest classifier documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] Fake News Detection Using RNN. https://www.kaggle.com/therealcyberlord/fake-news-detection-using-rnn