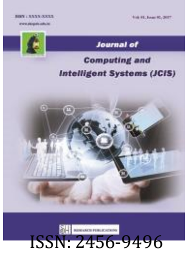
SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in



ISSN: 2456-9496

Fake News Detection Using Machine Learning Approaches

Poovizhi Magendiran#1, Jesmi Star#2

Received on 11 JUN 2022, Accepted on 14 JUN 2022

Abstract — Fake news is frequently spread on social media and other platforms, and it is a severe concern since it can cause considerable social and national harm and adverse effects. A great deal of research has previously been done on detection. This paper investigates the research on fake news detection and traditional machine learning models to determine which are the best in order to develop a product model with a supervised machine learning algorithm that can classify fake news as true or false, using Python Scikit-Learn and Natural Language Processing (NLP) for textual analysis. This procedure will result in feature extraction and vectorization. Then, the highest precision possible was obtained using feature selection methods to experiment and find the best fit features based on the confusion matrices results.

Keywords – Fake news detection, decision tree, logistic regression, machine learning, social media, information manipulation, natural language processing, textual investigation, NLTK random.

I. INTRODUCTION

In Fake news is comprised of factually incorrect information. This maintains a government's statistic deceit or inflated expenses of specific services for a country, leading to unrest in some countries, such as the Arab Spring. The House of Commons and the Crosscheck initiative, for example, are working to address issues such as author responsibility. However, their breadth is severely limited because they rely on human manual detection. Manual detection is neither accountable norfeasible in a world where millions ofitems are withdrawn or published every minute. Creating a system that creates a reliable automatic index scoring, or rating, for the reliability of multiple sources and news context could be one answer.

This paper demonstrates how to construct a model that can detect whether an article is real or fake based on its words, phrases, sources, and titles using supervised machine learning techniques on an annotated (labelled) dataset that has been manually categorized and validated. The feature selection methods are then used to explore and pick the best-fit featuresbased on the results of the confusionmatrix in order to get the highest precision. To construct the model, we suggest using a variety of categorization strategies. The product model will test anonymous data and display the results. Consequently, the product will be a model that detects and classifies fraudulent articles that can be used inthe future with any system.

2. PROBLEM STATEMENT

Using social media to consume news is a double-edged sword. People use social media to find and consume news because of the low cost, fast access, and rapid information delivery. On the other hand, it makes it easier to propagate "fake news," or low-quality, purposefully incorrect information. Fake news can harm people as well as society. As a result, detecting fake news on social media has become a prominent research focus in recent months.

Fake news is designed to persuade readers to trust false information, making it difficult and time-consuming to detect based on news content. As a result, users need to incorporate auxiliary data, such as social media behaviours, to help us make decisions. Second, exploiting this auxiliary data is challenging in and of itself, as consumers' social interactions with fake news generate massive, incomplete, unstructured, and noisy data.

*** Corresponding author: E-mail:** [1]poomca12@gmail.com, [2]jesmistar@gmail.com

[1]Assistant Professor, PG Department of computer science, Sacred Heart College (Autonomous), Tiruattur, Tamilnadu
[2]Research Scholar , PG Department of computer science, Sacred Heart College (Autonomous), Tirupattur, Tamilnadu

3. Related Work

3.1 Natural Language Processing

The term "natural language processing" is frequently used to refer to one or more system or algorithm specialities [7]. An algorithmic system's Natural Language Processing (NLP) grade allows for the integration of voice interpretation and speech production. It can also be used to recognize different languages' [5]. The activities used numerous language pipelines such as Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling to present a new ideal technique for extracting actions from English, Italian, and Dutch conversations. [6][8][9].

3.2 Machine Learning (ML) Classification

Machine Learning (ML) is a set of approaches that allows software systems to provide more accurate results without reprogramming. When creating predictions, data scientists specify the changes or features that the model must consider. After the training, the algorithm splits the learnt levels into new data [10]. To classify fake news, this research uses six algorithms.

3.3 Decision Tree

A flow chart-like arrangement Decision Tree helps to solve classification difficulties. [8] Internal nodes in the decision tree offer a condition or "test" on an attribute, with branchingdetermined by test conditions and results. When all characteristics have been Computed, the leaf node is given a class label. The categorization criterion is the distance between the root and the leaf. It is fantastic that it can function as a dependent and categorical variable. The decision tree algorithm excels in identifying and displaying the most significant components and their interactions. The algorithm is essential in creating new variables and features that can be used for data exploration and forecasting the target variable.

Tree-based learning algorithms are commonly used in predictive models that use supervised learning approaches. This approach is particularly good at illustrating non-linear relationships. It is also known as CART due to its ability to solve classification and regression problems.

3.5 Random Forest Classifier

Random Forests are based on the idea of producing many decision tree algorithms, each of which produces a unique outcome. The random forest is used to find outcomes that many decision trees have predicted. The random forest selects a subcategory of attributes from each group at random to generate variety in the decision trees. Regarding uncorrelated decision trees, Random Forest is the best method to utilize. When applied to similar trees, the overall result is the same as a single decision tree. Bootstrapping and feature randomness can be used to make uncorrelated decision trees.

3.6 Related Work on Fake News Detection

Decision Tree

The decision tree, which has a flow chart-like shape, is a valuable tool for overcoming classifying difficulties. The core nodes of the decision tree each offer a condition or "test" on an attribute, with branching based on the test conditions and outcomes. The leaf node is allocated a class label when all attributes are computed. The categorization criterion is the distance between the root and the leaf. This decision tree has a **99.79 percent** overall accuracy rate for detecting fake news. This approach aids in the classification of the dataset for false news detection, resulting in more precise results.

Random Forest

Random Forests are based on the idea of producing many decision tree algorithms, each of which produces a unique outcome. The random forest is used to find outcomes predicted by many decision trees. This algorithm has a 98.98 percent accuracy rating.

Logistic Regression

The Supervised Learning approach includes logistic regression, one of the most common Machine Learning algorithms. The outcome of a categorical dependent variable is predicted using logistic regression. This method has a **98.76 percent** accuracy rate in identifying fake news.

4. METHODOLOGY

The classification method is dealt with in the section. This method has been used to create a way of detecting fake news articles. In order to classify data, this method uses a neural network classifier. The first phases in solving this classification challenge include preprocessing, feature selection, collection training and testing, and classifier running.

Figure 1 illustrates the methodology of the fake news detector system. The process is based on doing a series of tests on a dataset using the decision tree, majority voting, and other classifier methods described in the previous section. Trials are conducted separately on each algorithm and in connection with others for the best accuracy and precision. The primary purpose is to use various classification methods to create a classification model that can be used as a guideline. A fake news scanner can be created by detecting details in the news and integrating the model into a Python application that can be used to discover bogus news data.

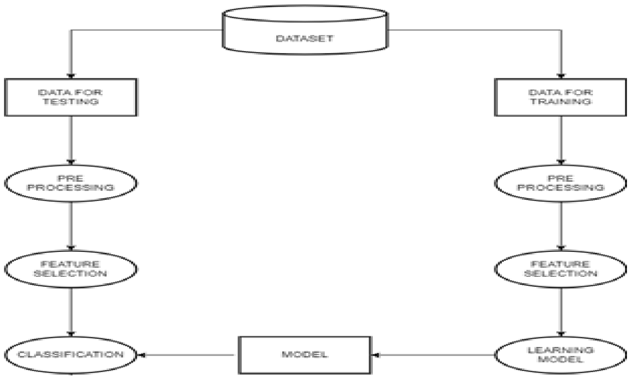


Figure 1: The methodology of the fake news detector system

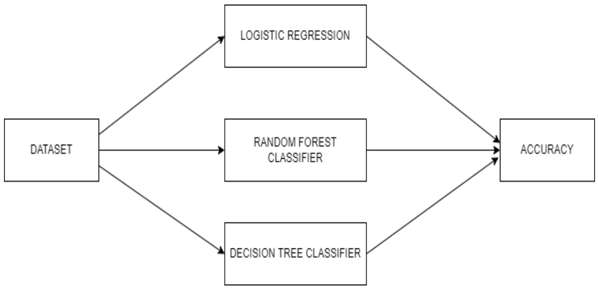


Figure 2: Different ML algorithms used to detect the fake news

The Python code has also been refactored to make it more efficient. Word clouds, decision trees, random forest classifiers, and logistic regression are employed as classification algorithms in this model. All of these algorithms strive for the most acceptable level of precision possible. Discordant and whenever possible, combine the average of them. The dataset is merged with several algorithms to identify fake news shown in figure 2. Before making a final judgment, the correctness of the study findings is evaluated.

4.1 NLTK (Natural Language Toolkit)

The (Natural Language Toolkit) is utilizedto implement POS and functionalities. Then, machine learning partition the dataset and create a classifier model (Nave Bays and Random Forest). The system successfully preprocesses the dataset after applying the NLTK, as shown in Figure 2, and generates a message for applying algorithms to the trained region. Themodel is developed using the response message after applying Naïve Bayes and Random forest to the system response. The precision is tested for authorization after testing and verifying the results on a test dataset. The model is then applied to previously unseen data using the data selected by the user.

The entire dataset is constructed, with half of the data being fictitious and the other half being genuine articles, giving the model a 50% reset accuracy. The fake and genuine datasetsare randomly selected for inclusion in our total dataset, with the remaining 20% as testing set after our system is completed. Users must first reduce noise by applying Stanford NLP for POS (Part of Speech) processing and tokenization of words and then encode the resulting data as integers and floating-point values so thatmachine learning approaches can use it. The python scikit-learn package, which includes essential functions like Count Vectorizer and Tiff Vectorizer, is used for tokenization and feature extraction of text data. A confusion matrix is used to visualize data in a graphical format, as shown in Figure 3.

The chosen dataset, FAKE NEWS DETECTION Master, and the data disinfecting and extraction techniques.

To determine the accuracy of the news, the following tests were performed.

1. The True-dataset has been preprocessed,whereas the Fake-dataset has not been Convert the given dataset to CSV format.

2. To reduce noise, use the NLP NLTK libraries and the scipy library.

3. Noise includes identifiers, dots, commas, and quotations, and the suffix is deleted by stemming terms. Using POS, the dataset will be turned into tokens and statistical values (Part of Speech). Select lexical features for feature extraction, such as word clouds, fake and absolute news frequency, data processing, and data modelling.

4. Use the Tfidf Vectorizer function in python Sklearn to extract unigram and bigram features. To create TF-IDF n-gram features, use this feature extraction library.

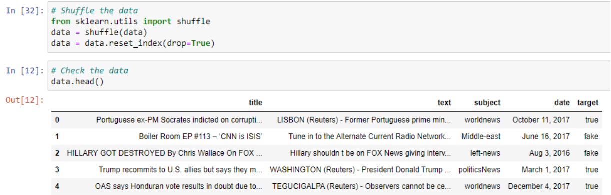
5. Split the dataset in half using pythonsklearn, with 70% train and 30% test.

6. Create a classification model ipynb fileafter completing all of the algorithms.

7. Use the model precision test part of the dataset to create a confusion matrix.

8. Check the clarity and correctness of fake and actual news.

9. Design an interface that will be used togauge how consumers react to previouslyunseen data.



```
In [32]: # Shuffle the data
         from sklearn.utils import shuffle
         data = shuffle(data)
         data = data.reset_index(drop=True)

In [12]: # Check the data
         data.head()

Out[12]:
```

	title	text	subject	date	target
0	Portuguese ex-PM Socrates indicted on corrupti...	LISBON (Reuters) - Former Portuguese prime min...	worldnews	October 11, 2017	true
1	Boiler Room EP #113 – 'CNN is ISIS'	Tune in to the Alternate Current Radio Network...	Middle-east	June 16, 2017	fake
2	HILLARY GOT DESTROYED By Chris Wallace On FOX ...	Hillary shouldn t be on FOX News giving interv...	left-news	Aug 3, 2016	fake
3	Trump recommits to U.S. allies but says they m...	WASHINGTON (Reuters) - President Donald Trump ...	politicsNews	March 1, 2017	true
4	OAS says Honduran vote results in doubt due to...	TEGUCIGALPA (Reuters) - Observers cannot be ce...	worldnews	December 4, 2017	true

Figure 4 shows how the data is selected at random and processed. There are two portions to the data: The first 75 percent of the data is training data, which aids the algorithm in distinguishing between correct and incorrect data. The data is labelled as 0 or 1, with 0 indicating false news and 1 indicating accurate information. After that, the remaining 25% of the data will be tested to see if it passes.

The algorithm % will be generated depending on the percentage of correct and incorrect responses, whether the news is authentic or fraudulent. Exploration of the data and the number of actual and phoney items in this fake news dataset are shown in figures 5,6,7 and 8.

5. RESULTS

This research work aims to place political news data from the Liar-dataset dataset, a New Benchmark Dataset for Fake News Detection divided into two categories: fake and trust news. On the "Liar" dataset was used in the research work. A confusion matrix represented the outcomes of the dataset analysis utilizing the three algorithms. The three detection techniques are listed below:

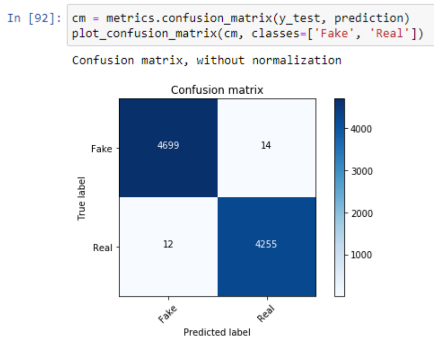
- Random Forest
- Decision Tree
- Logistic Regression

The confusion matrix is built automatically using Python code utilizing the cognitive learning module when the algorithm code isexecuted in the Anaconda platform.

The algorithms' Confusion Matrix is shown in Figures 9,10 and 11. Table 1 shows the Accuracy results of three algorithms.

Decision Tree

The categorization criterion is the distance between the root and the leaf. This decision tree has a 99.79 percent over all accuracy rate for detecting fake news.



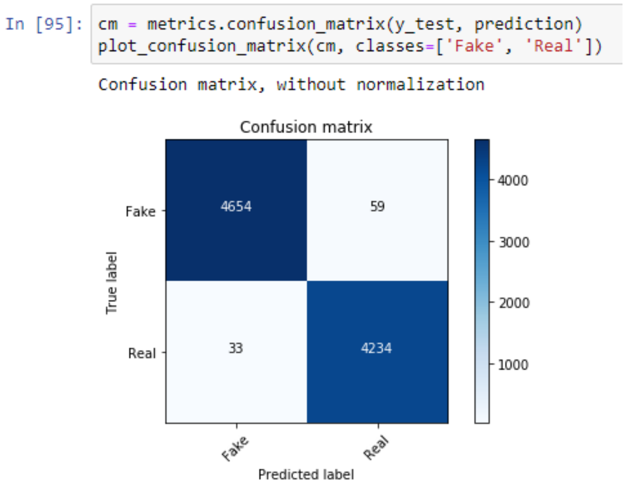
```
In [92]: cm = metrics.confusion_matrix(y_test, prediction)
         plot_confusion_matrix(cm, classes=['Fake', 'Real'])

Confusion matrix, without normalization
```

Figure 9: Confusion Matrix of Decision Tree

Random Forest

Random Forests are based on the idea of producing many decision tree algorithms, each of which produces a unique outcome. The random forest is used to find outcomes predicted by a large number of Decision trees. This algorithm has a **98.98 percent** accuracy rating.



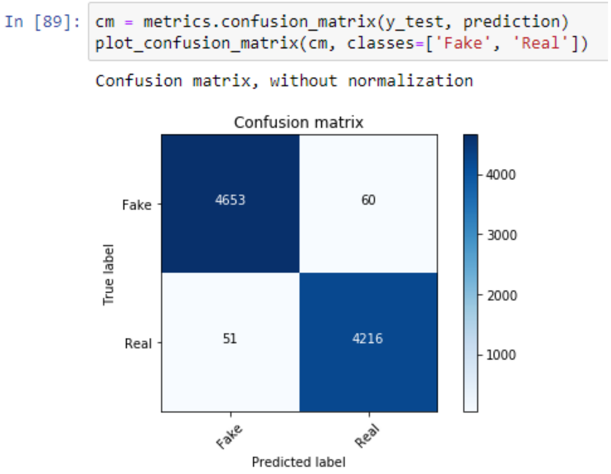
```
In [95]: cm = metrics.confusion_matrix(y_test, prediction)
         plot_confusion_matrix(cm, classes=['Fake', 'Real'])

Confusion matrix, without normalization
```

Figure 10: Confusion Matrix of Random Forest classifier

Logistic Regression:

The Supervised Learning approach includes logistic regression, one of the most common Machine Learning algorithms. The outcome of a categorical dependent variable is predicted using logistic regression. This method has a **98.76 percent** accuracy rate in identifying fake news.



```
In [89]: cm = metrics.confusion_matrix(y_test, prediction)
         plot_confusion_matrix(cm, classes=['Fake', 'Real'])

Confusion matrix, without normalization
```

Figure 11: Confusion Matrix of Logistic Regression

Table 1: Accuracy Resu lts of ML algorithms

1.	Logistic Regression	98.76%
2.	Random Forest Classifier	98.98%
3.	Decision Tree Classifier	99.71%

6. CONCLUSION

This study focuses on detecting bogus news using a two-stage evaluation process:categorization and disclosure. The first phase involves using social media to expose underlying concepts and principles of fake news. Several supervised learning algorithms are applied to investigate strategic approaches for detecting fake news during the discovery phase. Three approaches are typically used to detect fake news: Random Forest classifier, Decision Tree, and LogisticRegression. The Decision Tree Algorithm shows the highest level of precision **99.71 percent.**

REFERENCES

[1] Economic and Social Research Council. Using Social Mmedia. Available at: https://esrc.ukri.org/research/impact-toolkit/social-media/using-social-media

[2] Gil, P. Available at: https://www.lifewire.com/what-exactly-is-twitter-2483331. 2019, April 22.

[3] E. C. Tandoc Jr et al. "Defining fake news a typology of scholarly definitions". Digital Journalism , 1–17. 2017.

[4] J. Radianti et al. "An Overview of Public Concerns During the Recovery Period after a Major Earthquake: Nepal Twitter Analysis." HICSS '16 Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 136-145). Washington, DC, USA : IEEE. 2016.

[5] Alkhodair S A, Ding S H.H, Fung B C M, Liu J 2020 "Detecting breaking news rumors of emerging topics in social media" Inf. Process. Manag. 2020, 57, 102018.

[6] Jeonghee Yi et al. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. "In Data Mining, 2003. ICDM 2003. Third IEEE International Conference (pp. 427-434). http://citeseerx.ist.psu.edu. 200).2003

[7] Tapaswi et al. "Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit m sentences." Software Engineering (CONSEG), on Software Engineering (CONSEG), (pp. 1-4). IEEE. 2012

[8] Ranjan et al. "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi". In Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003). Semanticscholar. 2003

[9] MonaDiab et al. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of HLT-NAACL 2004: Short Papers (pp. 149–152). Boston, Massachusetts, USA: Association for Computational Linguistics. 2004

[10] Rouse, M. https://searchenterpriseai. techtarget.com/ definition/machine-learning-ML May 2018.

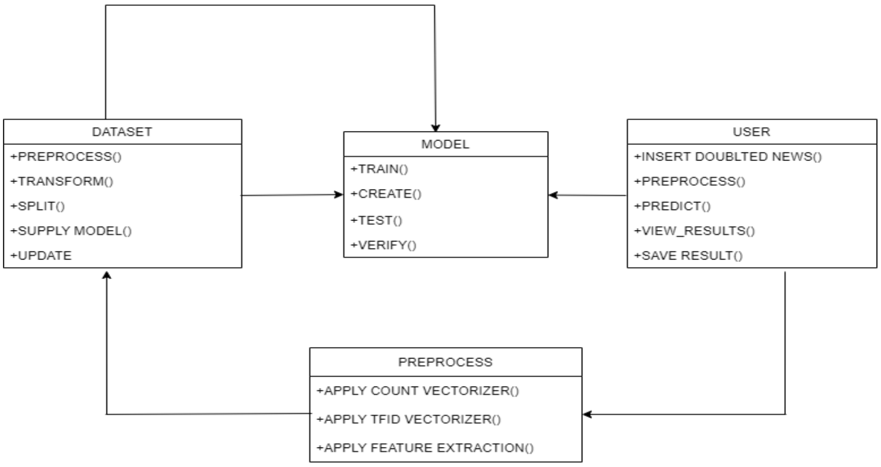
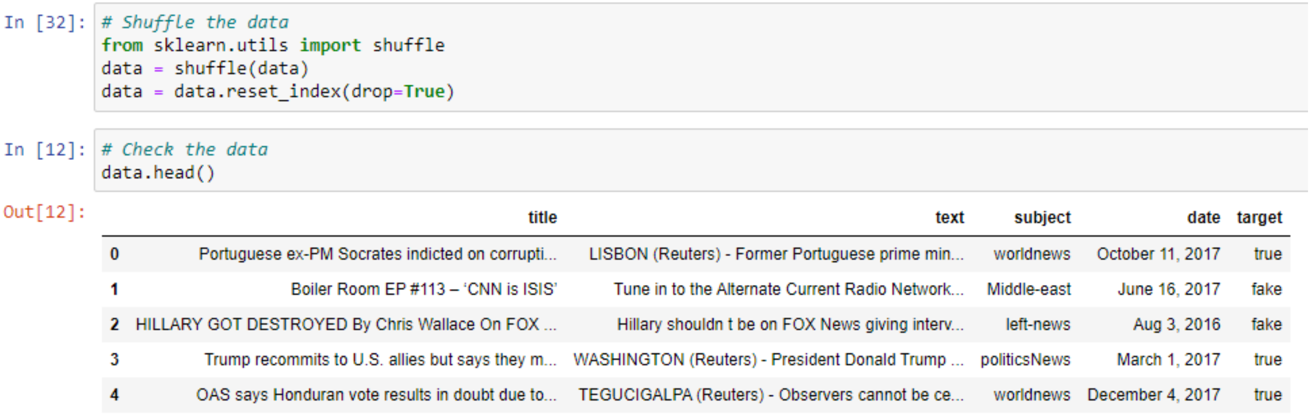


Figure 3: Model of a Fake Detector



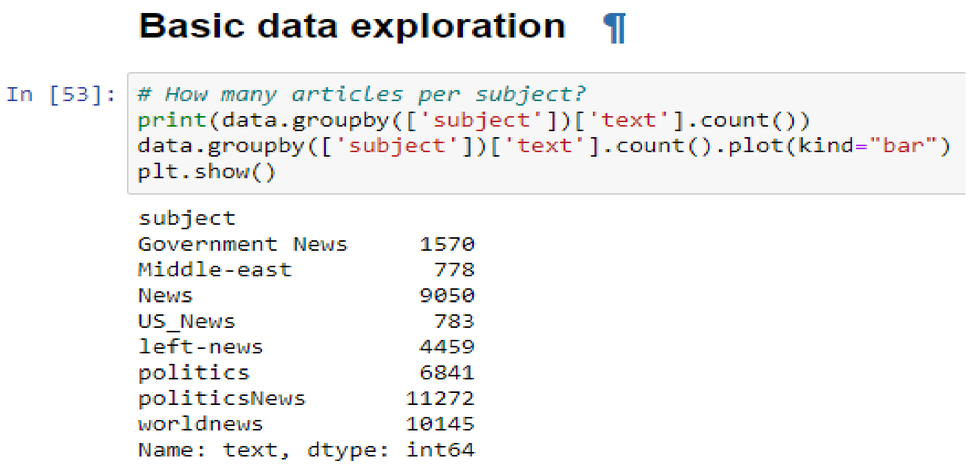
```
In [32]: # Shuffle the data
         from sklearn.utils import shuffle
         data = shuffle(data)
         data = data.reset_index(drop=True)

In [12]: # Check the data
         data.head()

Out[12]:
```

	title	text	subject	date	target
0	Portuguese ex-PM Socrates indicted on corrupti...	LISBON (Reuters) - Former Portuguese prime min...	worldnews	October 11, 2017	true
1	Boiler Room EP #113 – 'CNN is ISIS'	Tune in to the Alternate Current Radio Network...	Middle-east	June 16, 2017	fake
2	HILLARY GOT DESTROYED By Chris Wallace On FOX ...	Hillary shouldn t be on FOX News giving interv...	left-news	Aug 3, 2016	fake
3	Trump recommits to U.S. allies but says they m...	WASHINGTON (Reuters) - President Donald Trump ...	politicsNews	March 1, 2017	true
4	OAS says Honduran vote results in doubt due to...	TEGUCIGALPA (Reuters) - Observers cannot be ce...	worldnews	December 4, 2017	true

Figure 4: Processing the Dataset



Basic data exploration ¶

In [53]:

```
# How many articles per subject?
print(data.groupby(['subject'])['text'].count())
data.groupby(['subject'])['text'].count().plot(kind="bar")
plt.show()
```

```
subject
Government News      1570
Middle-east           778
News                 9050
US_News               783
left-news            4459
politics             6841
politicsNews        11272
worldnews           10145
Name: text, dtype: int64
```

Figure 5: Exploration into basic data

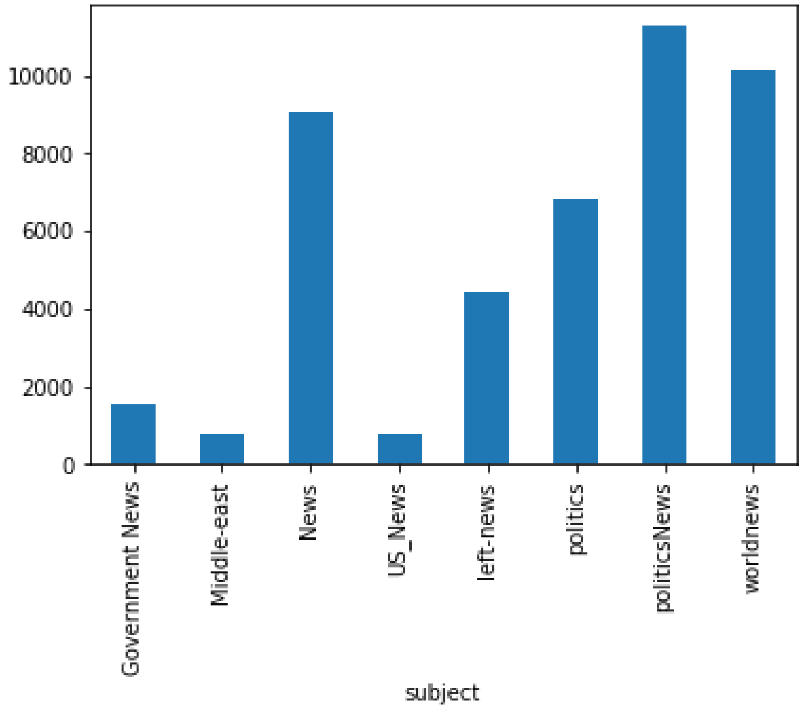
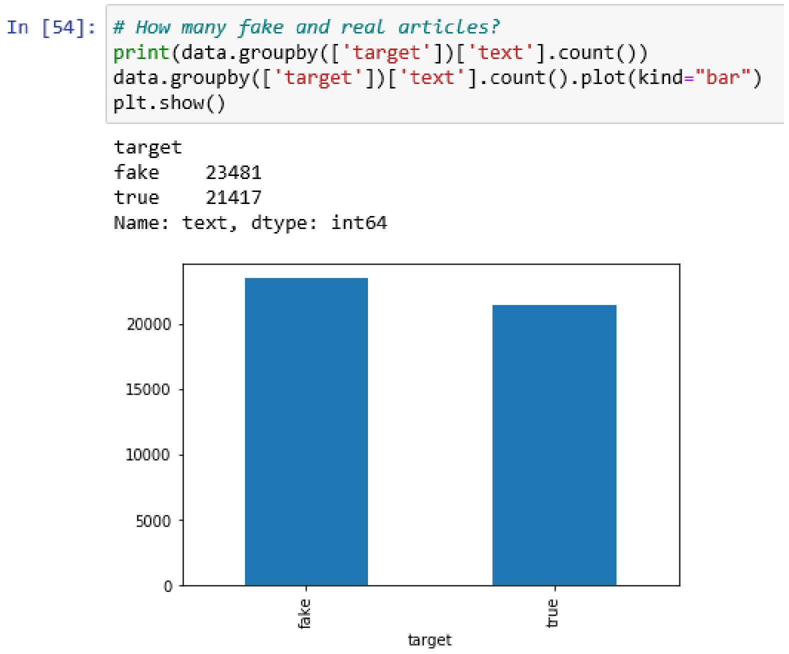


Figure 6: Exploration into basic data

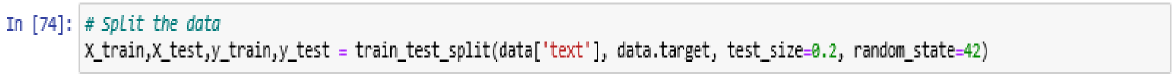


```
In [54]: # How many fake and real articles?
         print(data.groupby(['target'])['text'].count())
         data.groupby(['target'])['text'].count().plot(kind="bar")
         plt.show()

target
fake    23481
true    21417
Name: text, dtype: int64
```

Figure 7: Articles classified as fake and real

Peparing the data



```
In [74]: # Split the data
         X_train,X_test,y_train,y_test = train_test_split(data['text'], data.target, test_size=0.2, random_state=42)
```

Figure 8: Dataset after it has beed preprocessed