**UNIVERSITY INSTITUTE OF TECHNOLOGY**

**BARKATULLAH UNIVERSITY, BHOPAL**

Department Of Computer Science & Engineering

**MAJOR PROJECT (2020-2021)**

ON

**“FAKE NEWS DETECTION”**

Submitted for the partial fulfillment of the requirement for the

Award of Degree of

Bachelor of Engineering (B.E.)

Year 2020-21

By

**Rupali Patel Sandhya Rani Mourya Mansi Mittal Sonali Mahavar Girraj Yadav**

Under the Guidance

Of

**Mr. Amit Jha Dr.Divakar Singh** Asst. Professsor Head of Department CSE, UIT-BU CSE, UIT-BU

**UNIVERSITY INSTITUTE OF TECHNOLOGY**

**BARKATULLAH UNIVERSITY, BHOPAL**

Department Of Computer Science & Engineering

**CERTIFICATE**

**2020-21**

This is to certify that **Ms. Rupali Patel, Ms. Mansi Mittal, Ms. Sonali Mahavar, Ms. Sandhya Rani Mourya & Mr. Girraj** **Yadav** have successfully completed this project work titled **“Fake News Detection”** for the partial fulfillment of the award of Degree of “Bachelor Of Engineering” in Computer Science& Engineering in the year of 2020-21 from Barkatullah University Instiute Of Technology,Bhopal.

**Mr.Amit Jha Dr. Divakar Singh Dr. N.K. Gaur**

Project Guide Head Of Department Director

CSE UIT BU CSE UIT BU UIT BU

**UNIVERSITY INSTITUTE OF TECHNOLOGY**

**BARKATULLAH UNIVERSITY, BHOPAL**

Department Of Computer Science & Engineering

**DECLARATION**

**2020-2021**

We hereby declare that the work which is being presented in this report entitled, **“FAKE NEWS DETECTION”** submitted in the partial fulfillment of the requirement for the award of Degree of “Bachelor Of Engineering”in Computer Science and Engineering from University Institute Of Technology, BU Bhopal is an authentic record of our project work.

This is our original work and has not been submitted earlier for the award of any other degree,diploma or any other certificate.

**Rupali Patel Sandhya rani Mourya Girraj Yadav**

**Sonali Mahvar Mansi Mittal**

**ACKNOWLEDGEMENT**

It gives us immense pleasure to present this section as a tribute to those who always stood by us as a strong and acted as torchbearers for us.

Hereby, our first and foremost thanking goes to **Mr. Amit Jha** (Project Guide) for the knowledge and guidance provided to us on this project work.We gratefully thank him for extending to us her invaluable time and resources.

We would also like to thank **Dr. Divakar Singh,** Head of Dept(CSE), for providing effective plateform and support in the development of this project. And finally we would also like to render our thanks to our honorable Director **Dr. N.K. Gaur** for his support in this project, titled **“FAKE NEWS DETECTION”.**

Last but not least we would like to thank our parents and friends for their support and cooperation. and to those who indirectly have contributed in this project.

The project has indeed helped us to explore and implement more knowledgeable avenue related to our topic and for sure it will help us in our future.

**Submitted By:-**

Rupali Patel(R188237200033)

Mansi Mittal(R188237200024)

Sonali Mahavar(R188237200020)

Sandhya Rani Mourya(R188237200052)

Girraj Yadav(R188237200049)

**ABSTRACT**

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article.

In this work, we propose to use machine learning ensemble approach for automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, we train a combination of different machine learning algorithms using various ensemble methods and evaluate their performance on a real world dataset. Experimental evaluation confirms the superior performance of our proposed ensemble learner approach in comparison to individual learners.

**Keywords—Internet, Social Media, Fake News, Classification, Artificial Intelligence, Machine Learning.**

**TABLE OF CONTENT**

**TOPIC PAGE NO.**

LIST OF FIGURE I

LIST OF TABLES II

LIST OF ABBREVIATIONS III

**CHAPTER 1:INTRODUCTION 08**

**CHAPTER 2:REQUIREMENTS 11**

2.1 Software Requirements 11

2.2 Programming Language 11

2.3 Libraries 11

**CHAPTER 3: PROJECT DESCRIPTION 12**

3.1 Algorithms 12

3.2 Ensemble Learners 14

3.3 Benchmark Algorithms 16

3.4 Dataset 16

3.5 Performance Matrix 17

**CHAPTER 4: CODING AND OUTPUT SNAPSHOTS 19**

**CHAPTER 5: RESULT 29**

**CHAPTER 6: CONCLUSION 30**

**REFERENCES 31**

LIST OF TABLES

**S.No. Table No. Table Name Page No.**

1 Table 3.1 Confusion Matrix 16

2 Table 5.1 Accuracy 28

3 Table 5.2 Performance comparision 28

LIST OF ABBREVIATIONS

1.NLP Natural Language Processing

2.ML Machine Learning

3. DS Dataset

4. KNN K-Nearest Neighbour

5. SVM Support Vector Machine

6. LR Logistic Regression

7. DT Decision Tree

8. SGD Stochstic Gradient Descent

9. CNN Convolution Nueral Network

10.TF Term Frequency

11.IDF Inverse Document Frequency

12.MLP Multilayer Perceptron

13.CART Classification And Regression Trees

14.RF Random Forest

15.LSTM Long Short Term Memory Network

16.GBC Gradient Boosting Classifier

#### CHAPTER 1: INTRODUCTION

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. Besides other use cases, news outlets benefitted from the widespread use of social media platforms by providing updated news in near real time to its subscribers. The news media evolved from newspapers, tabloids, and magazines to a digital form such as online news platforms, blogs, social media feeds, and other digital media formats. It became easier for consumers to acquire the latest news at their fingertips. Facebook referrals account for 70% of traffic to news websites. These social media platforms in their current state are extremely powerful and useful for their ability to allow users to discuss and share ideas and debate over issues such as democracy, education, and health. However, such platforms are also used with a negative perspective by certain entities commonly for monetary gain and in other cases for creating biased opinions, manipulating mindsets, and spreading satire or absurdity. The phenomenon is commonly known as fake news.

There has been a rapid increase in the spread of fake news in the last decade, most prominently observed in the 2016 US elections. Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science. One such area affected by fake news is the financial markets, where a rumor can have disastrous consequences and may bring the market to a halt.

Our ability to take a decision relies mostly on the type of information we consume; our world view is shaped on the basis of information we digest. There is increasing evidence that consumers have reacted absurdly to news that later proved to be fake. One recent case is the spread of novel corona virus, where fake reports spread over the Internet about the origin, nature, and behavior of the virus. The situation worsened as more people read about the fake contents online. Identifying such news online is a daunting task.

Fortunately, there are a number of computational techniques that can be used to mark certain articles as fake on the basis of their textual content. Majority of these techniques use fact checking websites such as “PolitiFact” and “Snopes.” There are a number of repositories maintained by researchers that contain lists of websites that are identified as ambiguous and fake. However, the problem with these resources is that human expertise is required to identify articles/websites as fake. More importantly, the fact checking websites contain articles from particular domains such as politics and are not generalized to identify fake news articles from multiple domains such as entertainment, sports, and technology.

The World Wide Web contains data in diverse formats such as documents, videos, and audios. News published online in an unstructured format (such as news, articles, videos, and audios) is relatively difficult to detect and classify as this strictly requires human expertise. However, computational techniques such as natural language processing (NLP) can be used to detect anomalies that separate a text article that is deceptive in nature from articles that are based on facts. Other techniques involve the analysis of propagation of fake news in contrast with real news. More specifically, the approach analyzes how a fake news article propagates differently on a network relative to a true article. The response that an article gets can be differentiated at a theoretical level to classify the article as real or fake. A more hybrid approach can also be used to analyze the social response of an article along with exploring the textual features to examine whether an article is deceptive in nature or not.

A number of studies have primarily focused on detection and classification of fake news on social media platforms such as Facebook and Twitter. At conceptual level, fake news has been classified into different types; the knowledge is then expanded to generalize machine learning (ML) models for multiple domains. The study by Ahmed et al. included extracting linguistic features such as n-grams from textual articles and training multiple ML models including K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), linear support vector machine (LSVM), decision tree (DT), and stochastic gradient descent (SGD), achieving the highest accuracy (92%) with SVM and logistic regression. According to the research, as the number of  increased in -grams calculated for a particular article, the overall accuracy decreased. The phenomenon has been observed for learning models that are used for classification. Shu et al. achieved better accuracies with different models by combining textual features with auxiliary information such as user social engagements on social media. The authors also discussed the social and psychological theories and how they can be used to detect false information online. Further, the authors discussed different data mining algorithms for model constructions and techniques shared for features extraction. These models are based on knowledge such as writing style, and social context such as stance and propagation.

A different approach is followed by Wang. The author used textual features and metadata for training various ML models. The author focused mainly on using convolutional neural network (CNN). A convolutional layer is used to capture the dependency between the metadata vectors, followed by a bidirectional LSTM layer. The max-pooled text representations were concatenated with the metadata representation from the bidirectional LSTM, which was fed to fully connected layer with a softmax activation function to generate the final prediction. The research is conducted on a dataset from political domain which contains statements from two different parties. Along with that, some metadata such as subject, speaker, job, state, party, context, and history are also included as a feature set. Accuracy of 27.7% was achieved with combination of features such as text and speaker, whereas 27.4% accuracy was achieved by combining all the different metadata elements with text. A competitive solution is provided by Riedel et al, which is a stance detection system that assigns one of four labels to an article, “agree,” “disagree,” “discuss,” or “unrelated,” depending on the conformity of article headline with article text. The authors used linguistic properties of text such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF) as a feature set, and a multilayer perceptron (MLP) classifier is used with one hidden layer and a softmax function on the output of the final layer. The dataset contained articles with a headline, body, and label. The system’s accuracy on the “disagree” label on test examples was poor, whereas it performs best with respect to the “agree” label. The authors used a simple MLP with some fine-tuned hyperparameters to achieve an overall accuracy of 88.46%. Shu et al. also discussed several varieties of veracity assessment methods to detect fake news online. Two major categories of assessment methods are explored: one is linguistic cue approaches and the other is network analyses approaches. A combination of both creates a more robust hybrid approach for fake news detection online. Linguistic approaches involve deep syntax, rhetorical structure, and discourse analysis. These linguistic approaches are used to train classifiers such as SVM or naïve Bayes models. Network-based approaches included analyzing and processing social network behavior and linked data. A unique approach is followed by Vosoughi et al.to explore the properties of news spread on social media; i.e., the authors discussed the spread of news (rumors) on social media such as Twitter and analyzed how the spread of fake news differs from real news in terms of its diffusion on Twitter. Multiple analysis techniques are discussed in the paper to explore the spread of fake news online, such as the depth, the size, the maximum breadth, the structural virality, the mean breadth of true and false rumor cascades at various depths, the number of unique Twitter users reached at any depth, and the number of minutes it takes for true and false rumor cascades to reach depth and number of Twitter users.

Characteristics of Fake News:

They often have grammatical mistakes. They are often emotionally colored. They often try to affect reader’s opinion on some topics. Their content is not always true. They often use attention seeking words and news format and click baits. They are too good to be true. Their sources are not genuine most of the times.

## CHAPTER 2: REQUIREMENTS

2.1.Software

* Jupyter Notebook

2.2.Programming Language

* Python

2.3.Libraries

* Numpy
* Pandas
* Matplotlib
* sklearn

CHAPTER 3: PROJECT DESCRIPTION

There are numerous reputed websites that post legitimate news contents, and a few other websites such as PolitiFact and Snopes which are used for fact checking. In addition, there are open repositories which are maintained by researchers to keep an up-to-date list of currently available datasets and hyperlinks to potential fact checking sites that may help in countering false news spread. However, we selected three datasets for our experiments which contain news from multiple domains (such as politics, entertainment, technology, and sports) and contain a mix of both truthful and fake articles.

The input features are then used to train the different machine learning models. Each dataset is divided into training and testing set with a 70/30 split, respectively. The articles are shuffled to ensure a fair allocation of fake and true articles in training and tests instances.

The learning algorithms are trained with different hyperparameters to achieve maximum accuracy for a given dataset, with an optimal balance between variance and bias. Each model is trained multiple times with a set of different parameters using a grid search to optimize the model for the best outcome. Using a grid search to find the best parameters is computationally expensive; however, the measure is taken to ensure the models do not overfit or underfit the data.

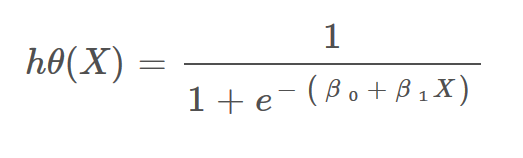
various ensemble techniques such as bagging, boosting, and voting classifier are explored to evaluate the performance over the multiple datasets. We used two different voting classifiers composed of three learning models: the first voting classifier is an ensemble of logistic regression, random forest, and KNN, whereas the second voting classifier consists of logistic regression, linear SVM, and classification and regression trees (CART). The criteria used for training the voting classifiers is to train individual models with the best parameters and then test the model based on the selection of the output label on the basis of major votes by all three models. We have trained a bagging ensemble consisting of 100 decision trees, whereas two boosting ensemble algorithms are used, XGBoost and AdaBoost. A *k*-fold (*k* = 10) cross validation model is employed for all ensemble learners. The learning models used are described in detail in Section 3.1. To evaluate the performance of each model, we used accuracy, precision, recall, and F1 score metrics as discussed in Section 3.5.

##### 3.1. Algorithms

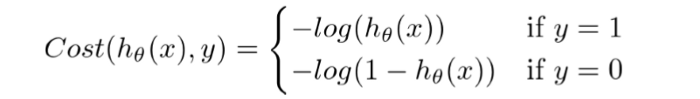
We used the following learning algorithms in conjunction with our proposed methodology to evaluate the performance of fake news detection classifiers.

###### **3.1.1. Logistic Regression**

As we are classifying text on the basis of a wide feature set, with a binary output (true/false or true article/fake article), a logistic regression (LR) model is used, since it provides the intuitive equation to classify problems into binary or multiple classes . We performed hyperparameters tuning to get the best result for all individual datasets, while multiple parameters are tested before acquiring the maximum accuracies from LR model. Mathematically, the logistic regression hypothesis function can be defined as follows:



Logistic regression uses a sigmoid function to transform the output to a probability value; the objective is to minimize the cost function to achieve an optimal probability. The cost function is calculated as shown:



###### **3.1.2. Support Vector Machine**

Support vector machine (SVM) is another model for binary classification problem and is available in various kernels functions. The objective of an SVM model is to estimate a hyperplane (or decision boundary) on the basis of feature set to classify data points [29]. The dimension of hyperplane varies according to the number of features. As there could be multiple possibilities for a hyperplane to exist in an *N*-dimensional space, the task is to identify the plane that separates the data points of two classes with maximum margin.

Kernels are usually used to fit data points that cannot be easily separable or data points that are multidimensional.

###### **3.1.3. Multilayer Perceptron**

A multilayer perceptron (MLP) is an artificial neural network, with an input layer, one or more hidden layers, and an output layer. MLP can be as simple as having each of the three layers; however, in our experiments we have fine-tuned the model with various parameters and number of layers to generate an optimum predicting model.

###### **3.1.4. K-Nearest Neighbors (KNN)**

KNN is an unsupervised machine learning model where a dependent variable is not required to predict the outcome on a specific data. We provide enough training data to the model and let it decide to which particular neighborhood a data point belongs. KNN model estimates the distance of a new data point to its nearest neighbors, and the value of *K* estimates the majority of its neighbors’ votes; if the value of *K* is 1, then the new data point is assigned to a class which has the nearest distance. The mathematical formulae to estimate the distance between two points can be calculated as follows :

##### KNN.png

##### 3.2. Ensemble Learners

We proposed using existing ensemble techniques along with textual characteristics as feature input to improve the overall accuracy for the purpose of classification between a truthful and a false article. Ensemble learners tend to have higher accuracies, as more than one model is trained using a particular technique to reduce the overall error rate and improve the performance of the model. The intuition behind the ensemble modeling is synonymous to the one we are already used to in our daily life such as requesting opinions of multiple experts before taking a particular decision in order to minimize the chance of a bad decision or an undesirable outcome. For example, a classification algorithm can be trained on a particular dataset with a unique set of parameters that can produce a decision boundary which fits the data to some extent. The outcome of that particular algorithm depends not only on the parameters that were provided to train the model, but also on the type of training data. If the training data contains less variance or uniform data, then the model might overfit and produce biased results over unseen data. Therefore, approaches like cross validation are used to minimize the risk of overfitting. A number of models can be trained on different set of parameters to create multiple decision boundaries on randomly chosen data points as training data. Hence, using ensemble learning techniques, these problems can be addressed and mitigated by training multiple algorithms, and their results can be combined for near optimum outcome. One such technique is using voting classifiers where the final classification depends on the major votes provided by all algorithms. However, there are other ensemble techniques as well that can be used in different scenarios such as the following.

###### **3.2.1. Random Forest (RF)**

Random forest (RF) is an advanced form of decision trees (DT) which is also a supervised learning model. RF consists of large number of decision trees working individually to predict an outcome of a class where the final prediction is based on a class that received majority votes. The error rate is low in random forest as compared to other models, due to low correlation among trees . Our random forest model was trained using different parameters; i.e., different numbers of estimators were used in a grid search to produce the best model that can predict the outcome with high accuracy. There are multiple algorithms to decide a split in a decision tree based on the problem of regression or classification. For the classification problem, we have used the Gini index as a cost function to estimate a split in the dataset. The Gini index is calculated by subtracting the sum of the squared probabilities of each class from one. The mathematical formula to calculate the Gini index () is as follows:

Gind=1-Summation(Pi^2)

**3.2.2. Bagging Ensemble Classifiers**

bootstrap aggregating, or in short bagging classifier, is an early ensemble method mainly used to reduce the variance (overfitting) over a training set. Random forest model is one of the most frequently used as a variant of bagging classifier. Intuitively, for a classification problem, the bagging model selects the class on the basis of major votes estimated by  number of trees to reduce the overall variance, while the data for each tree is selected using random sampling with replacement from overall dataset. For regression problems, however, the bagging model averages over multiple estimates.

###### **3.2.3. Boosting Ensemble Classifiers**

boosting is another widely used ensemble method to train weak models to become strong learners. For that purpose, a forest of randomized trees is trained, and the final prediction is based on the majority vote outcome from each tree. This method allows weak learners to correctly classify data points in an incremental approach that are usually misclassified. Initially equal weighted coefficients are used for all data points to classify a given problem. In the successive rounds, the weighted coefficients are decreased for data points that are correctly classified and are increased for data points that are misclassified. Each subsequent tree formed in each round learns to reduce the errors from the preceding round and to increase the overall accuracy by correctly classifying data points that were misclassified in previous rounds. One major problem with boosting ensemble is that it might overfit to the training data which may lead to incorrect predictions for unseen instances. There are multiple boosting algorithms available that can be used for both the purposes of classification and regression. In our experiments we used XGBoost and AdaBoost algorithms for classification purpose.

###### **3.2.4. Voting Ensemble Classifiers**

voting ensemble is generally used for classification problems as it allows the combination of two or more learning models trained on the whole dataset. Each model predicts an outcome for a sample data point which is considered a “vote” in favor of the class that the model has predicted. Once each model predicts the outcome, the final prediction is based on the majority vote for a specific class. Voting ensemble, as compared to bagging and boosting algorithms, is simpler in terms of implementation. As discussed, bagging algorithms create multiple subsets of data by random sampling and replacement from the whole dataset, thus creating a number of datasets. Each dataset is then used to train a model, while the final result is an aggregation of outcome from each model. In case of boosting, multiple models are trained in a sequential manner where each model learns from the previous by increasing weights for the misclassified points, thus creating a generic model that is able to correctly classify the problem. However, voting ensemble on the other hand is a combination of multiple independent models that produces classification results that contribute to the overall prediction by majority voting.

##### 3.3. Benchmark Algorithms

In this section, we discuss the benchmark algorithms with which we compare the performance of our methodology.

###### **3.3.1. Linear SVM**

We use linear SVM approach proposed in. To ensure a meaningful comparison, we trained the linear SVM on the feature set as discussed in with 5-fold cross validation. Note that the approach is referred to as Perez-LSVM in the text.

###### **3.3.2. Convolutional Neural Network**

Wang used convolutional neural network (CNN) for automatic detection of fake news. We employed the same approach using our dataset. However, we could not use the feature set of Wang as the dataset contains only short statements. The approach is referred to as Wang-CNN in the text.

###### **3.3.3. Bidirectional Long Short-Term Memory Networks**

Wang also used bidirectional long short-term memory networks (Bi-LSTM), and we used the same approach with different feature sets. The approach is referred to as Wang-Bi-LSTM in the text.

##### 3.4. Dataset

The datasets we used in this study are open source and freely available online. The data includes both fake and truthful news articles from multiple domains. The truthful news articles published contain true description of real world events, while the fake news websites contain claims that are not aligned with facts. The conformity of claims from the politics domain for many of those articles can be manually checked with fact checking websites such as politifact.com and snopes.com

The dataset is called the “ISOT Fake News Dataset” (hereafter referred to as DS) which contains both true and fake articles extracted from the World Wide Web. The true articles are extracted from reuters.com which is a renowned news website, while the fake articles were extracted from multiple sources, mostly websites which are flagged by politifact.com. The dataset contains a total of 44,898 articles, out of which 21,417 are truthful articles and 23,481 fake articles. The total corpora contain articles from different domains, but most prominently target political news.

##### 3.5. Performance Metrics

To evaluate the performance of algorithms, we used confusion metrix.

Confusion Matrix: A confusion matrix is a table that is often used to describe the performance of a classification model (or ―classifier‖) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Table 3.1: Confusion Matrix

|  |  |  |
| --- | --- | --- |
| TOTAL | Class 1 (Predicted) | Class 1 (Predicted) |
| Class 1  (Actual) | TP | FN |
| Class 2  (Actual) | FP | TN |

**3.5.1 Accuracy**

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used:

Accuracy = (TP+TN)%(TP+TN+FP+FN)

In most cases, high accuracy value represents a good model, but considering the fact that we are training a classification model in our case, an article that was predicted as true while it was actually false (false positive) can have negative consequences; similarly, if an article was predicted as false while it contained factual data, this can create trust issues. Therefore, we have used three other metrics that take into account the incorrectly classified observation, i.e., precision, recall, and F1-score.

***3.5.2. Recall***

Recall represents the total number of positive classifications out of true class. In our case, it represents the number of articles predicted as true out of the total number of true articles.

Recall=TP% (TP+FN)

###### **3.5.3. Precision**

Conversely, precision score represents the ratio of true positives to all events predicted as true. In our case, precision shows the number of articles that are marked as true out of all the positively predicted (true) articles:

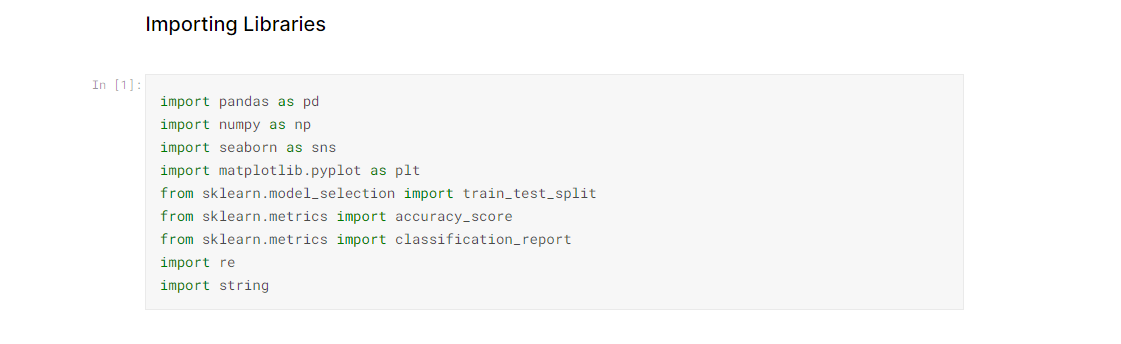
Precision=TP% (TP+FP)

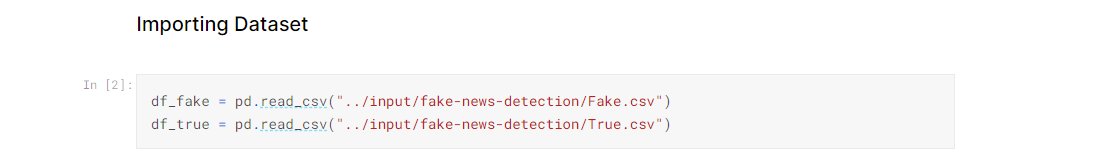
###### **3.5.4. F1-Score**

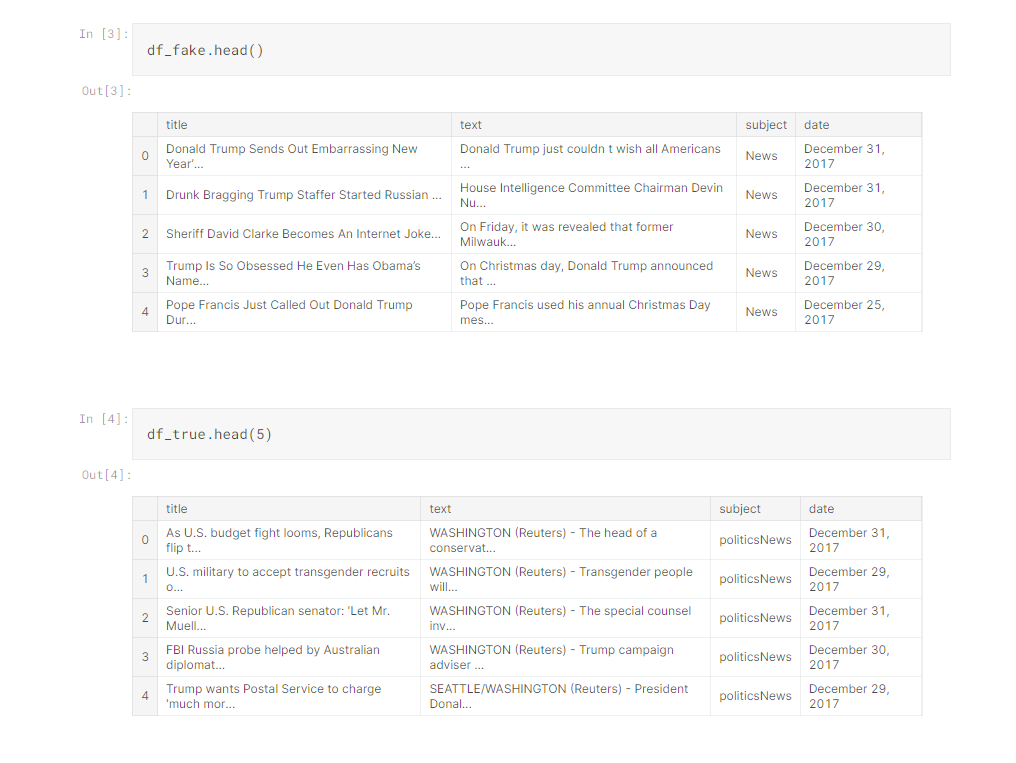
F1-score represents the trade-off between precision and recall. It calculates the harmonic mean between each of the two. Thus, it takes both the false positive and the false negative observations into account. F1-score can be calculated using the following formula:

F1-Score=2(Precision \* Recall)%(Precision + Recall)

#### CHAPTER 4: CODING AND OUTPUT SNAPSHOTS

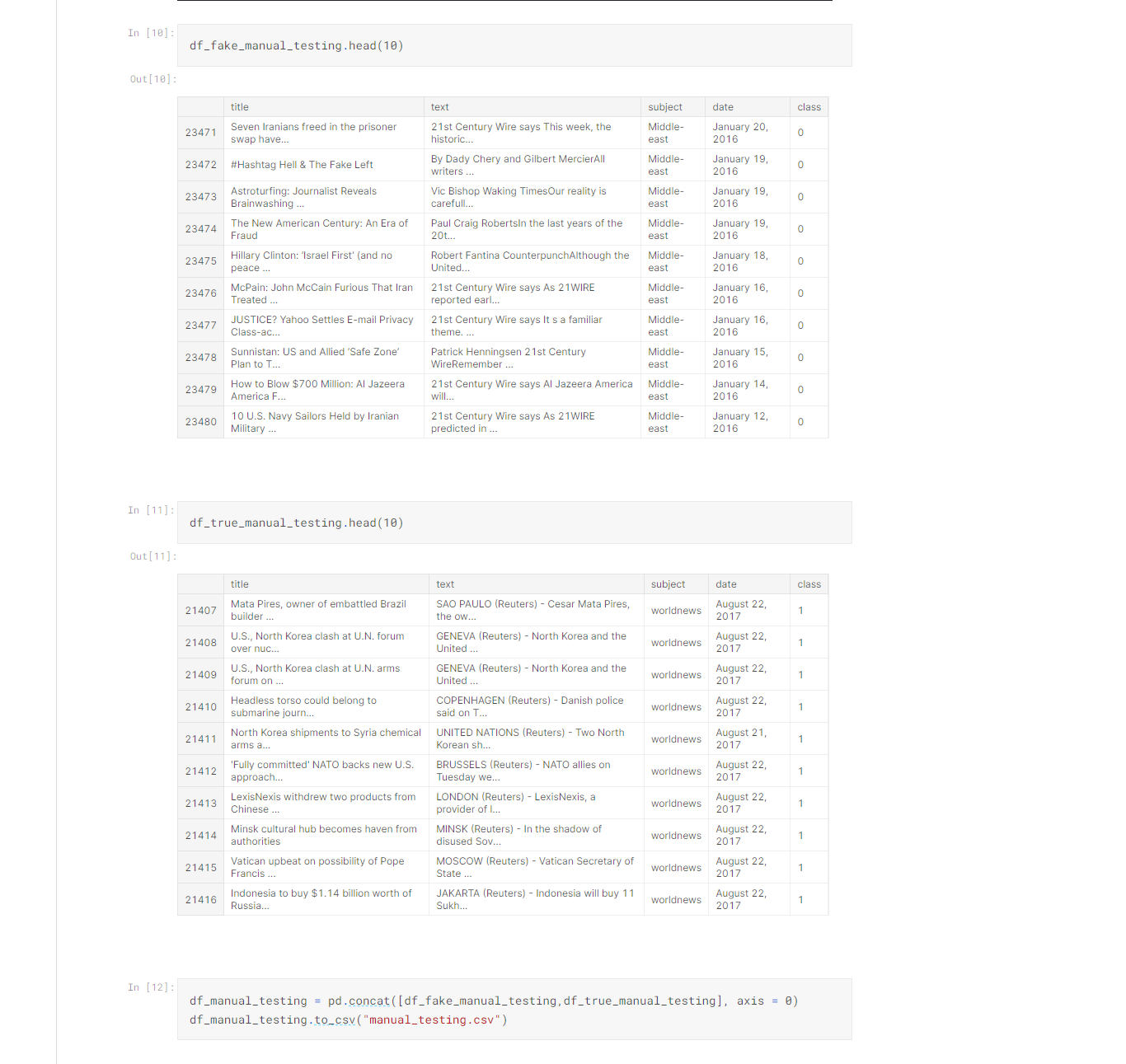


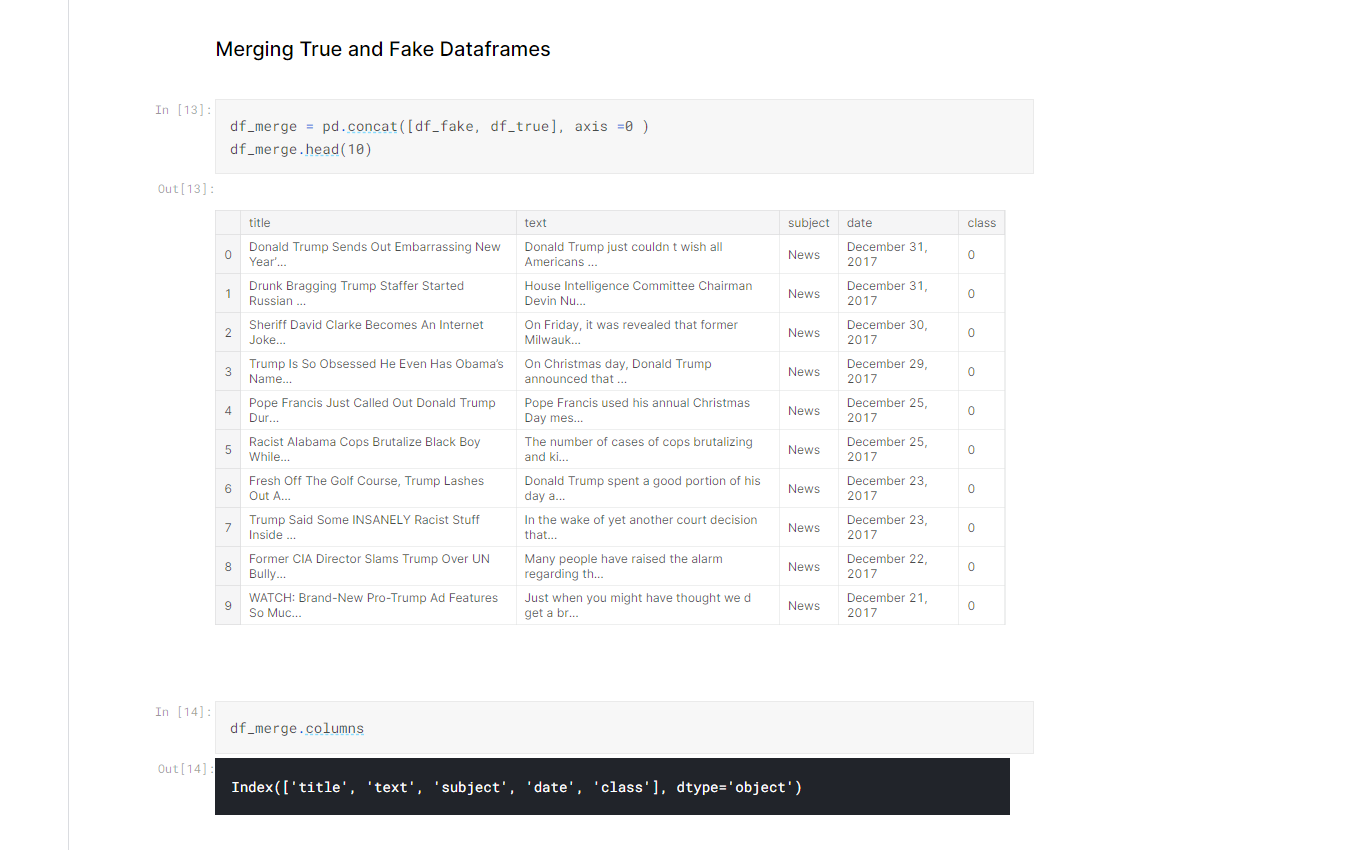


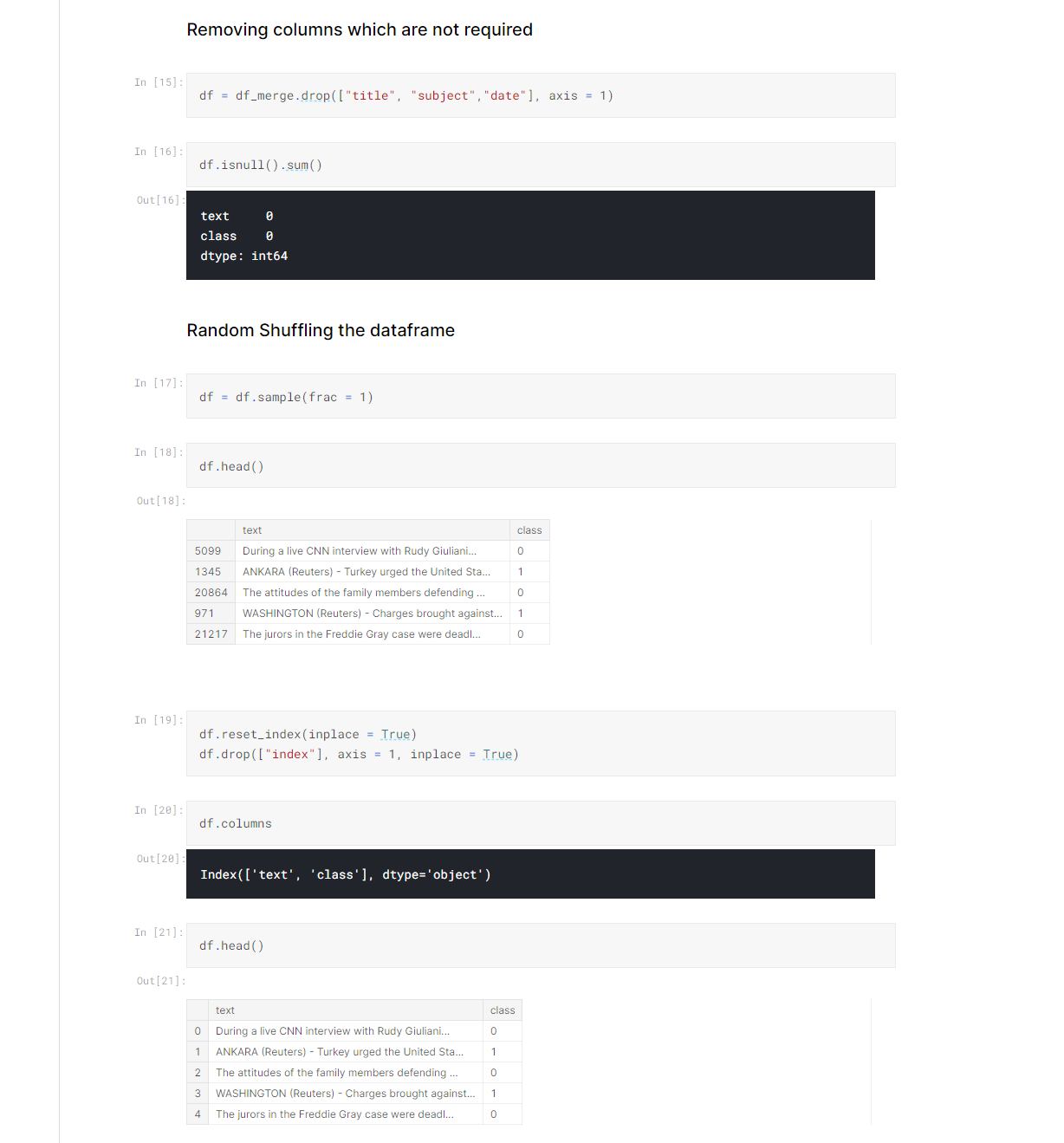


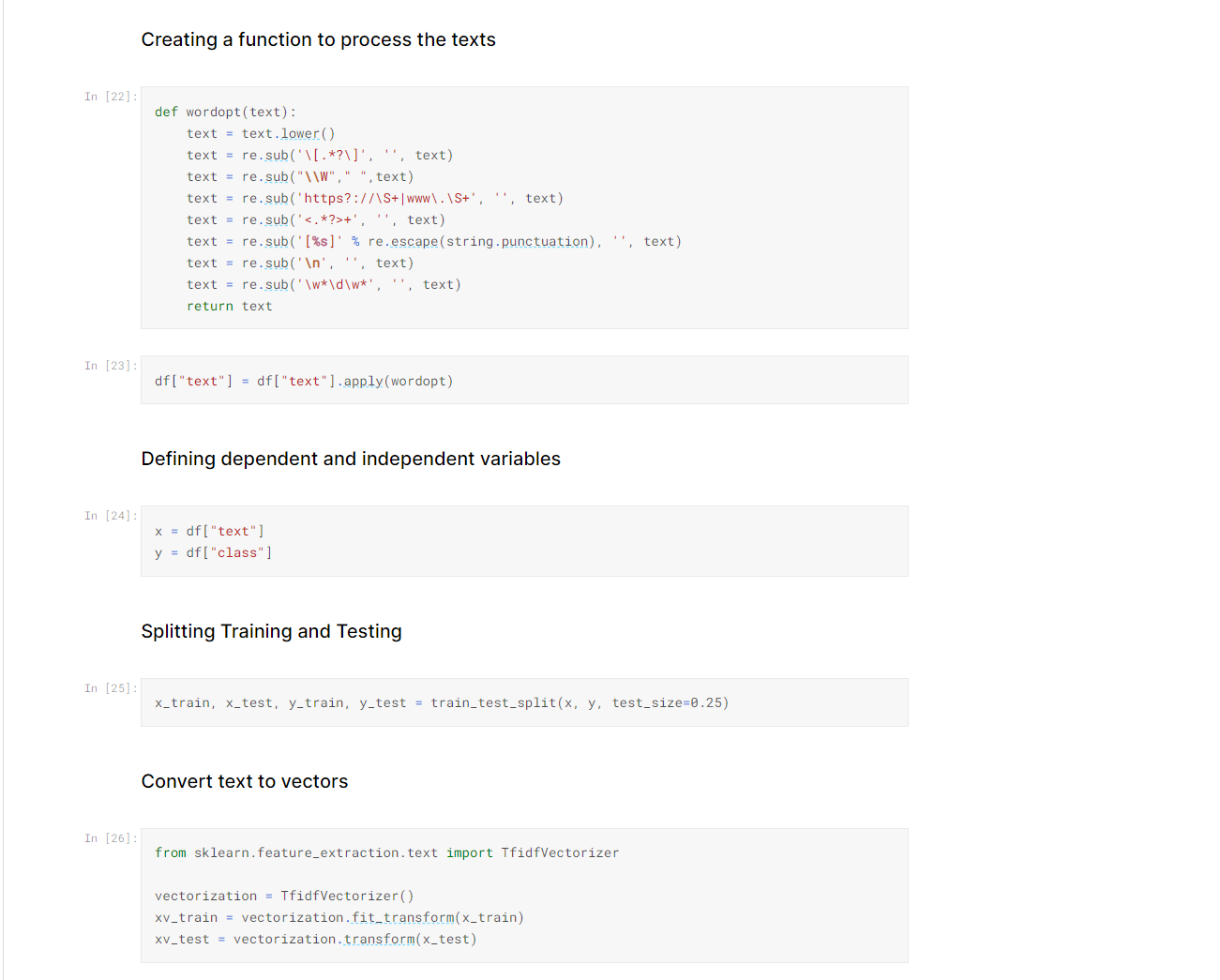


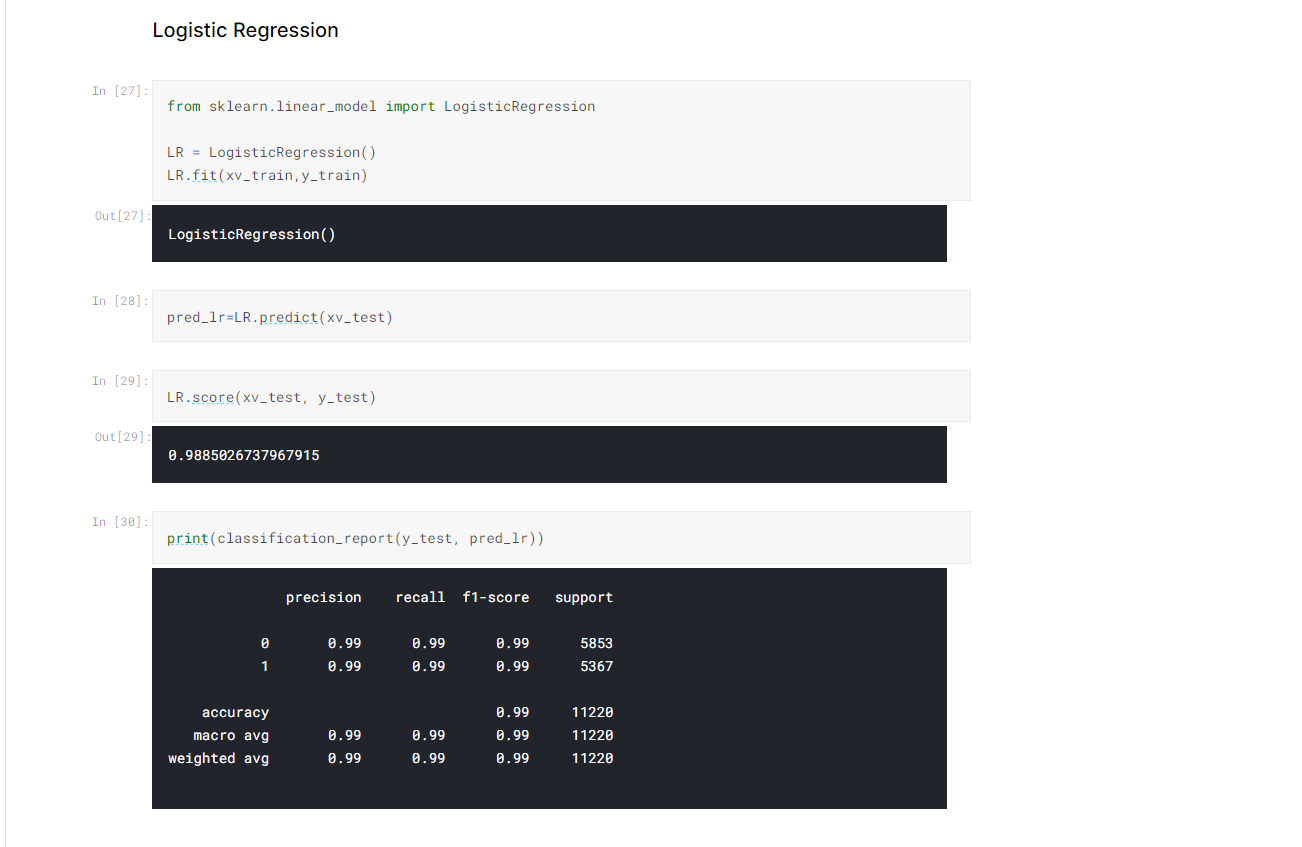


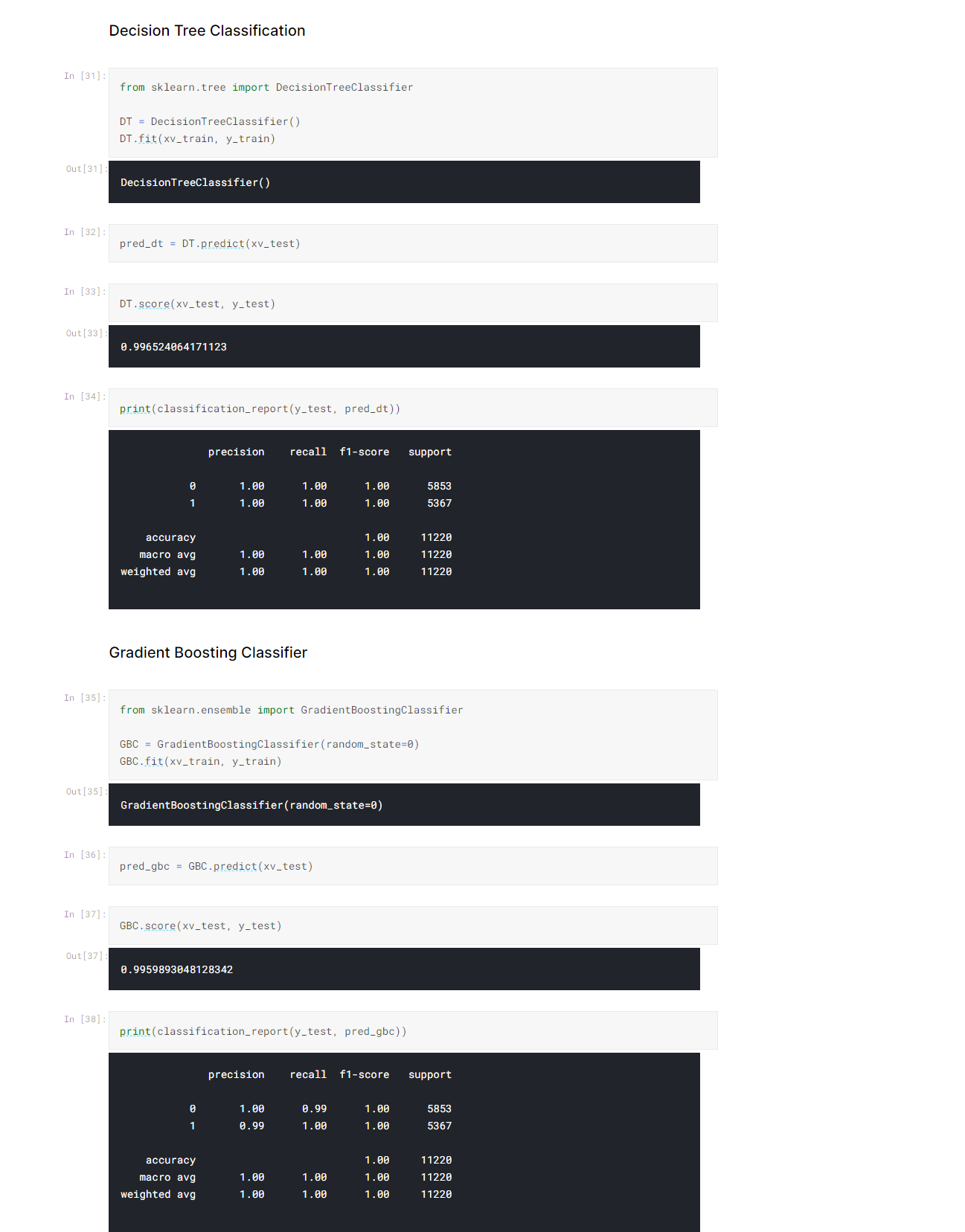


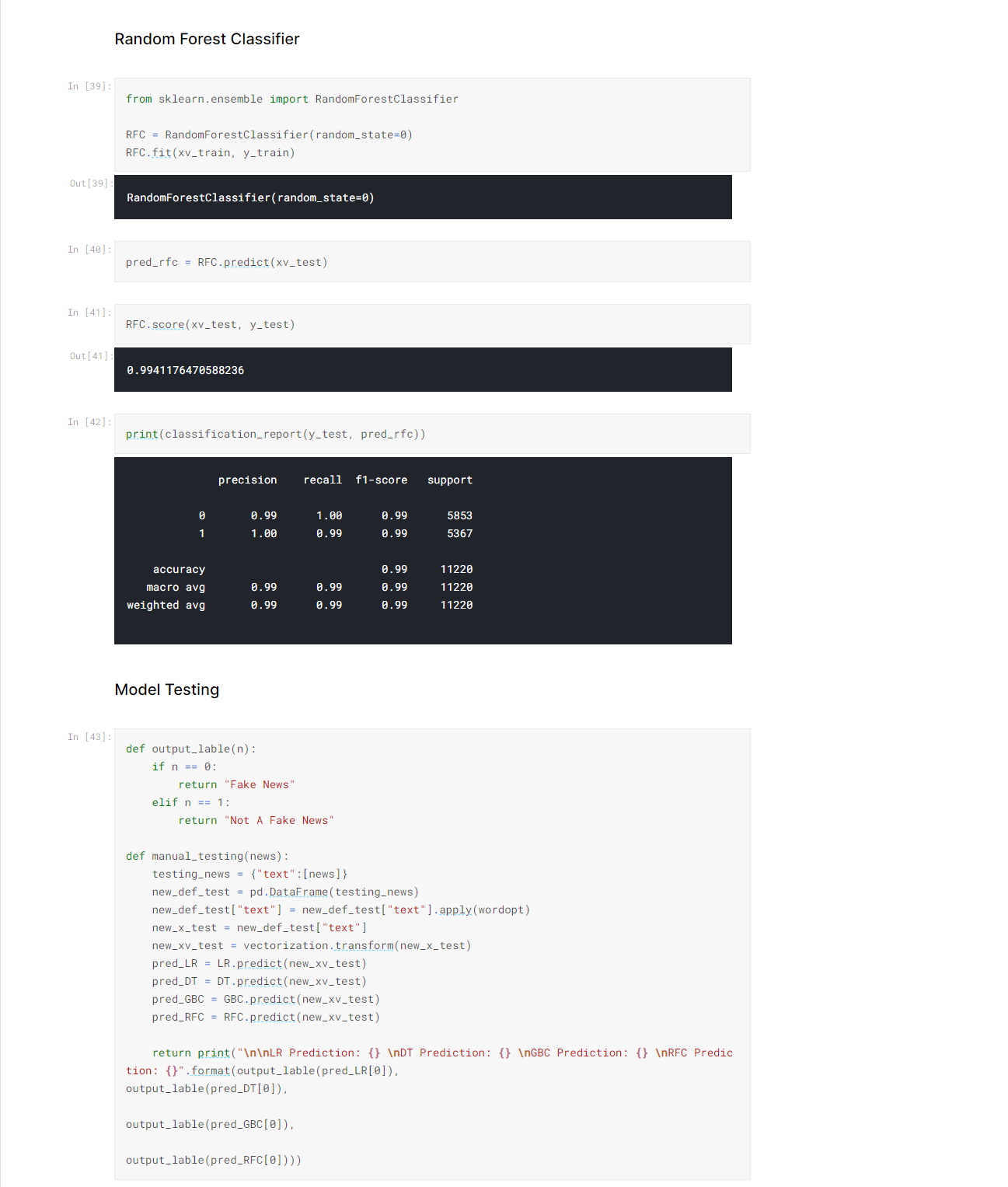


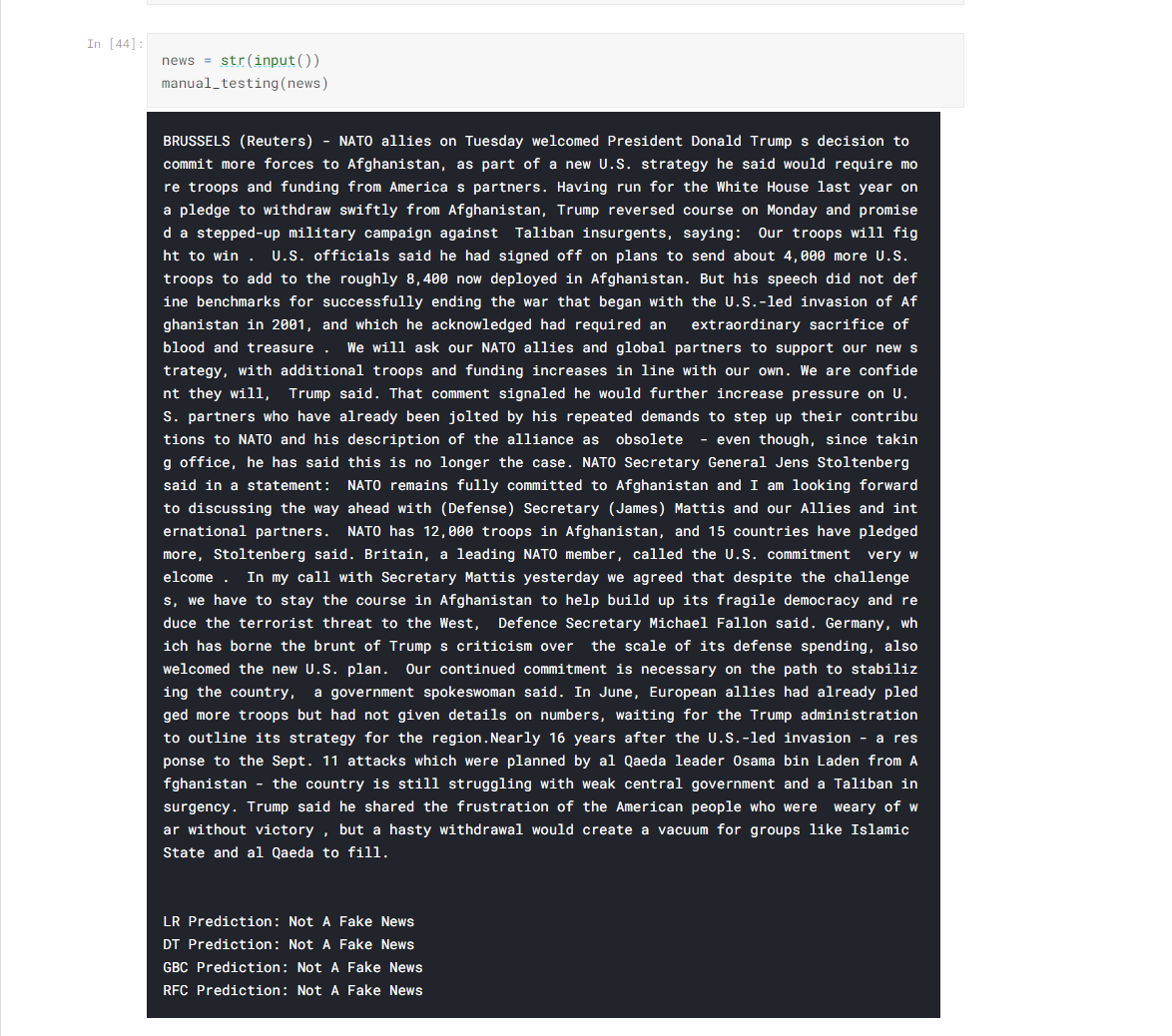


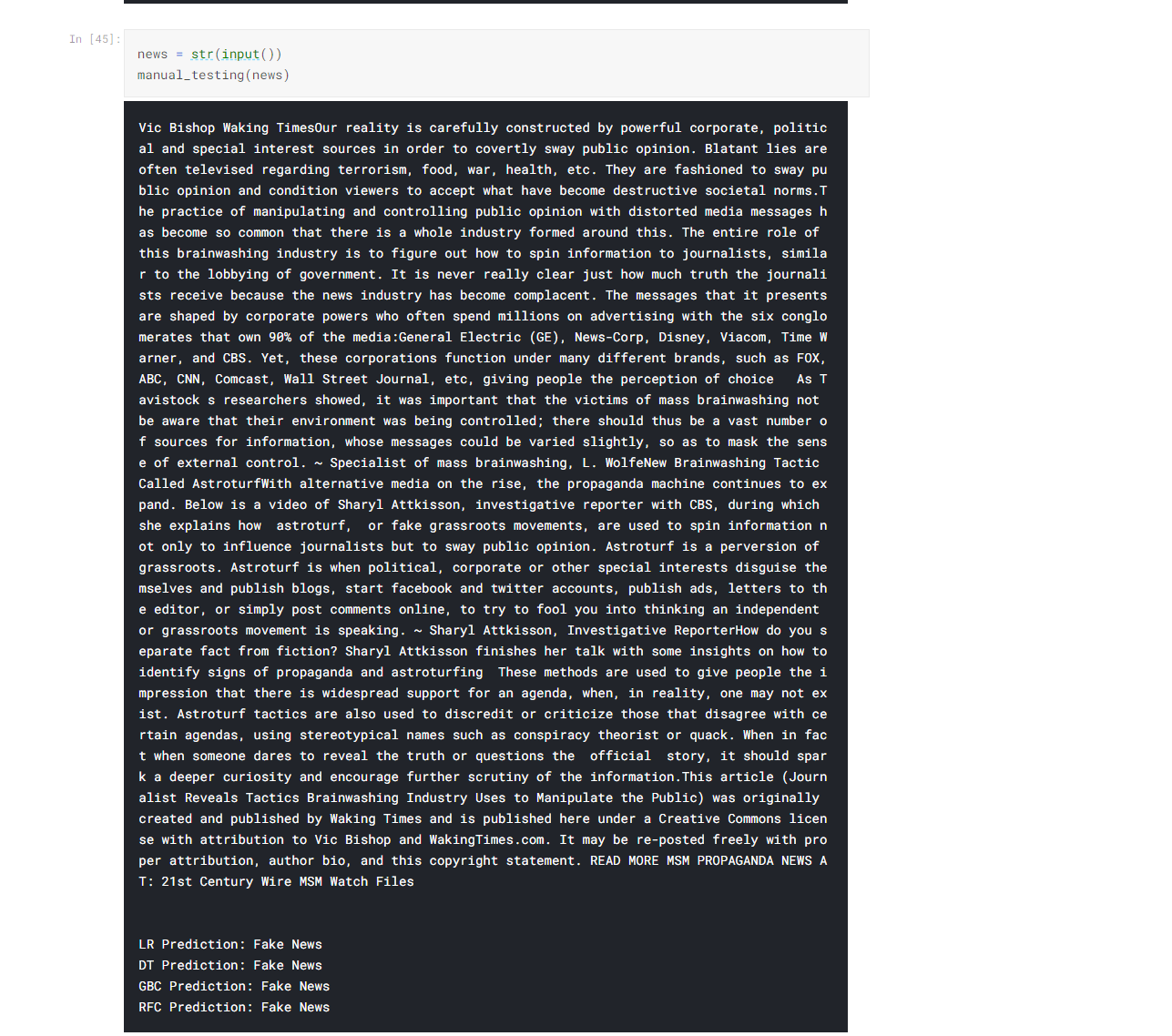












#### CHAPTER 5: RESULT AND DISCUSSION

#### Table 5.1 summarizes the accuracy achieved by each algorithm on the four considered datasets. It is evident that the maximum accuracy achieved on DS (ISOT Fake News Dataset) is 99.6%, achieved by decision tree algorithm. Logistic regression classifiers achieved an accuracy of 98% . gradient boosting classifier achieved 99.5%. and random forest achieved 99.4%.

#### Table 5.1:Accuracy

|  |  |
| --- | --- |
| ALGORITHM | ACCURACY |
| LR | 98% |
| RF | 99.4% |
| DT | 99.6% |
| GBC | 99.5% |

#### However, accuracy score alone is not a good measure to evaluate the performance of a model; therefore, we also evaluate performance of learning models on the basis of recall, precision, and F1-score.

#### Table 5.2:Performance comparision

|  |  |  |  |
| --- | --- | --- | --- |
| ALGORITHM | RECALL | PRECISION | F1 SCORE |
|  | 0.99 | 0.99 | 0.99 |
| Random Forest | 0.99 | 1.0 | 0.99 |
| Decision tree | 1.0 | 1.0 | 1.0 |
| Gradient Boosting | 1.0 | 0.99 | 1.0 |

#### From above observation we can say that on our dataset, random forest and decision performed best on the basis of precion and gradient boosting and decision tree performs good on the basis of recall and f1 score.

#### So overall we can found that decision tree performed best in our model.

#### CHAPTER 6: CONCLUSION

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text.The data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The primary aim is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles using an LIWC tool and used the feature set as an input to the models. The learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others. We used multiple performance metrics to compare the results for each algorithm. The ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Graph theory and machine learning techniques can be employed to identify the key sources involved in spread of fake news. Likewise, real time fake news identification in videos can be another possible future direction.

#### REFERENCES

1. A. D. Holan, *2016 Lie of the Year: Fake News*, Politifact, Washington, DC, USA, 2016.
2. D. M. J. Lazer, M. A. Baum, Y. Benkler et al., “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.View at: [Publisher Site](https://doi.org/10.1126/science.aao2998) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20science%20of%20fake%20news&author=D.%20M.%20J.%20Lazer&author=M.%20A.%20Baum&author=Y.%20Benkler%20et%20al.&publication_year=2018)
3. H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.View at: [Publisher Site](https://doi.org/10.1257/jep.31.2.211) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Social%20media%20and%20fake%20news%20in%20the%202016%20election&author=H.%20Allcott%20&author=M.%20Gentzkow&publication_year=2017)
4. <https://data-flair.training/>
5. https://www.kaggle.com/datasets/