Bachelor Thesis
Computer Science
05 2022

Comparison of Random Forest and Gradient Boosting algorithms for detecting fake news articles on media

Veera Venkata Subrahmanaya Varma Mudunuri
Manish Chowdary Boppana

Dept. Computer Science & Engineering
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

This thesis is submitted to the Department of Computer Science & Engineering at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to 10 weeks of full-time studies.

Contact Information:
Author(s):
Varma Mudunuri
E-mail: vemu20@student.bth.se
Manish Chowdary Boppana
E-mail: mady20@student.bth.se

University advisor:
PhD Mahwish Anwar
Dept. Computer Science

Dept. Computer Science & Engineering
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se/didd
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Context. Detection of fake news articles on Social media is really very important. People believe in those fake news articles and they trust that the news is true. So, it is very essential to classify fake news and real news articles on media and bring awareness among common people. Detection of fake news articles on media is important. Media is common and easily accessible which is why news articles shared through media is accepted by majority people. It is essential to classify the trustworthiness of news articles to control incorrect news articles propagation. This idea led to the research on fake news articles detection on social media from data sets taken from Scopus and Kaggle and using supervised machine learning algorithms.
Objectives. The main objective of this research is to detect fake news articles on media with machine learning algorithms namely Random Forest and Gradient Boosting. The detection of fake news articles is performed in the data sets taken from Scopus and Kaggle data bases. These data sets are a combination of both real and fake news articles.
Methods. Many have done research on detecting fake news articles on media. In this paper we use public data sets from Kaggle and Scopus, and apply the chosen algorithms Random Forest and Gradient Boosting to identify fake news articles on media. We evaluate the results using Accuracy, Recall, Precision and F1 score evaluation metrics.
Results.Among the two selected supervised machine learning algorithms, our research found that Gradient Boosting algorithm has better Accuracy and better F1 scores to detect fake news articles on given data sets.
Conclusions. We conclude that Gradient Boosting algorithm has better Accuracy compared with Random Forest algorithm and Gradient Boosting algorithm is efficient in detecting fake news articles on media.

Keywords: Gradient Boosting, Machine learning, Random Forest, TF-IDF vectorization.

List of Figures

1.1 Random Forest Classifier 6
1.2 Gradient Boosting Classifier 6

3.1 Work flow 10
3.2 Code 11
3.3 Code 12
3.4 Code 13

4.1 Table with evaluation metrics Data Set1 15
4.2 Table with evaluation metrics Data Set2 15
4.3 Table with evaluation Data Set 1 and 2 15
4.4 Bar Representation of Data Set1 16
4.5 Bar Representation of Data Set2 16
4.6 Bar Representation of Mixed Data Set 17

Contents

Abstract **i**

1 Introduction **1**
1.1 Machine Learning 2
1.2 Train and Test Split Function 3
1.3 Label Encoding 4
1.4 Aims and Objectives 4
1.5 Research Question 5
1.6 Background 5

2 Related Work **7**

3 Method **9**
3.1 Evaluation Metrics: 9
3.2 Experimental Design 10
3.3 Data Set Explanation and Processing 11

4 Results **14**

5 Discussion **18**

6 Conclusions and Future Work **19**

References **20**

Chapter 1

Introduction

Media platforms are a source of information for the people. Sometimes people use this media to access the latest news and for different reasons they share the information to other people [10]. The news articles delivered on the internet may be true or false. This delivery and propagation of fake news articles results in several problems. This issue of fake news articles is creating differences among people, creating disputes between two groups of people and sometimes leading to rifts between nations. Therefore, it is highly important to detect the fake news articles on media platforms to bring awareness among people about its truthfulness or fakeness. [2].

People are highly exposed to news articles on media media on regular basis. These fake news articles make people lose trust in the news articles published on media. Detection of fake news articles is challenging[3], but machine learning can help [13] with the appropriate detection techniques. Machine learning and deep learning are considered for fake news articles detection as it is cheap and fast. Machine Learning is the scientific study of computational algorithms and statistical models designed to compete with human intelligence. These computational algorithms use past information available to the learner to improve performance and perform accurate predictions. In Machine learning, the system learns from past data and apply that knowledge to predict future outcomes. Machine Learning is being used in every field, centred upon how their behaviour objects are classified and predicted future outcomes. Some detect fake news by comparing Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree and neural networks based on Accuracy. [5].

In this thesis we use Random Forest classifier and Gradient boosting classifier to detect fake news articles on media. We have taken two different data sets one from Scopus and the other from Kaggle data bases. These data sets are a combination of both fake news and real news articles. Each data set has approximately 20000 articles in them. Both the algorithms differ in their own way [15]. The biggest reason for choosing the Random Forest algorithm is that it provides better results for both categorical values and continuous data. It is also flexible to both

classification and regression. Gradient Boosting algorithms can optimize different functions which makes it flexible, and it often provides predictive accuracy. [11].

1.1 Machine Learning

Machine Learning is the field of science which deals with programming computers that take information known as data and learn from that data [9]. Machine Learning is a subset of Artificial Intelligence, which imitates human intelligence into machines, these machines think in the way of humans and mimic their behaviour. Any device's attributes that matches human brain can be applied with the term Machine Learning, such as critical thinking, problem-solving, learning, decision-making, etc. In recent years, many sectors like government, health care, E-commerce, marketing, and sales have been using Machine learning [4].

According to the labeled data, methods in Machine learning can be classified into three categories: supervised learning method, semi-supervised learning method and unsupervised learning method.

- Supervised Learning methods predict target variables in the presence of labeled data in which this method is trained to predict target variables based on input data. A supervised learning model helps predict outcomes of unseen data by learning from labeled training data [18]. This learning model is categorized into two categories. A predictive modelling technique is defined as regression when the target variable is numerical. A predictive modelling technique is defined as Classification when the target variable is categorical.

- The unsupervised learning method does not have output data to its corresponding input data. Unlike the supervised learning method here, data is unlabeled [6]. Unsupervised learning methods aids to discover unknown patterns in data. But the task complexity is more and unpredictable when compared to the supervised learning model. Commonly used unsupervised techniques are clustering, neural networks, and detection etc.

- The semi-Supervised learning model uses both supervised and unsupervised learning methods. In a semi-supervised learning model, during training of the small data amount of labeled data is combined with a large amount of unlabeled data [21]. This learning model discovers patterns in the unlabeled data learnt from a small amount of labelled data and uses this for predictions.

1.2 Train and Test Split Function

"Sklearn" or "scikit-learn" is a python library that consists of various data processing features, which are used for clustering, model selection, and classification. "Model-selection" is a method to frame a blueprint. This blueprint is used to analyze data and then using that data to measure new data. While making predictions, the selection of appropriate models aids to generate accurate results. In order to test a model using specific data set, first that particular model must be trained against the same data set. To split one data set and perform training and testing, the "train-test-split" function can be used. In evaluating data mining models, the separation of data into training and testing is an important part, where the majority of data is used for training, and a small amount of data is used for the testing purposes of the model [17]. "train-test-split" is a function in Sklearn model selection for splitting a dataset into two subsets. One is for training, and another is for testing the model. This function automatically divides data into subsets instead to perform that operation manually. By default, train-test-split takes random partitions to split into two subsets of the data. This procedure is widely used as it is fast and easy to perform. This procedure is not recommended in cases where there is a small dataset, additional configurations are needed, and the dataset is not balanced.

- The basic syntax of train-test-split would be like: train-test-split(X, y train-size=0., test-size=0., random-state=*)
- There are several parameters in test-train-split, they are:
  - X, y – In general, "X" is the variable used to store the dataset for training purposes, "y" is the variable used to store the dataset for testing purposes.train-size – This parameter is used to set the size of the dataset for training purposes, "None" is used as a default argument, "int" is used when an exact number of samples is known, "float" is used which ranges from 0.1 to 1.0.
  - test-size – This parameter is used to set the size of the dataset for testing purposes, "None" is used as a default argument, "int" is used when an absolute number of samples is known, "float" is used which ranges from 0.1 to 1.0, if train-size is also assigned "None" then 0.25 is set to complement test-size.
  - random-state – This parameter is used to control the shuffling during the splitting of the data. "None" is used as the default argument, "int" is used to reproduce output across multiple function calls.
- By default, the train-test-split function splits the data into random subsets.

- The standard way to split is 80:20 for training and testing respectively, adjustments are recommended depending on parameter complexity and size of the dataset.

1.3 Label Encoding

In Machine Learning, usually, datasets are dealt with datasets that contain labels with one or more columns. These labels are represented in either words or numbers. To be easily interpreted by humans and make the data understandable, words are used to label training data. Label Encoding means converting labels from numeric form to machine-readable form. Then the better way of operating those labels will be decided by Machine Learning algorithms. Label Encoding is a vital data pre processing step in supervised learning model. Label encoder is a utility class that aids in normalizing labels into numeric values between 0 to (number of classes)-1 1 . For example, there is a column in data height, and it consists of three labels: tall, medium, short. After performing label encoding, the labels in the column height can be converted into 0, 1, 2. Where 0 represents tall, 1 represents medium, and 2 illustrates short. There is a small limitation for label encoding that is, label encoding converts data into the machine-understandable way. As it assigns unique numbers starting from zero to each class of a dataset, this leads to priority issues while training datasets. For example, a label with a high value can be taken as a high priority than with a lower value.

1.4 Aims and Objectives

The aim of this thesis is to find out the degree of originality of a news article on media by developing a machine learning model using Random Forest and Gradient Boosting classifiers and later testing the model manually. The detection of fake news articles is performed on two different data sets taken from Kaggle and Scopus databases.
The objectives of this thesis are:
1. To create a machine learning model that can efficiently detect the authenticity of a news article on media using Random Forest and Gradient Boosting algorithms.
2. To train and classify the two models.
3. To compare the results of the two models.

1.5 Research Question

RQ: Which of the machine learning algorithms Random Forest or Gradient Boosting Classifier is efficient in detecting fake news articles on media on data sets taken from Kaggle and Scopus?

1.6 Background

Random Forest Classifier is flexible, easy to use machine learning algorithm which gives good results. This algorithm can be used for both classification and regression tasks. Generally, Random Forest algorithm builds multiple decision trees and merges them to get more perfect and accurate prediction [16].

The two major differences between Random Forest and Gradient Boosting algorithms varies in building trees and combining results. Random forest algorithm builds each tree independently while gradient boosting algorithm builds one tree at a time. While in terms of combining results random forest combines results at the end of process, gradient boosting combines results along the process [20].

Classifiers: In this paper we have used random forest classifier and gradient boosting classifier. Random Forest classifier is a tree based calculation. Here several number of multiple decision trees are build and later they all are merged to produce more accurate results. The following figure shows how random forest classifier predicts the result [14].

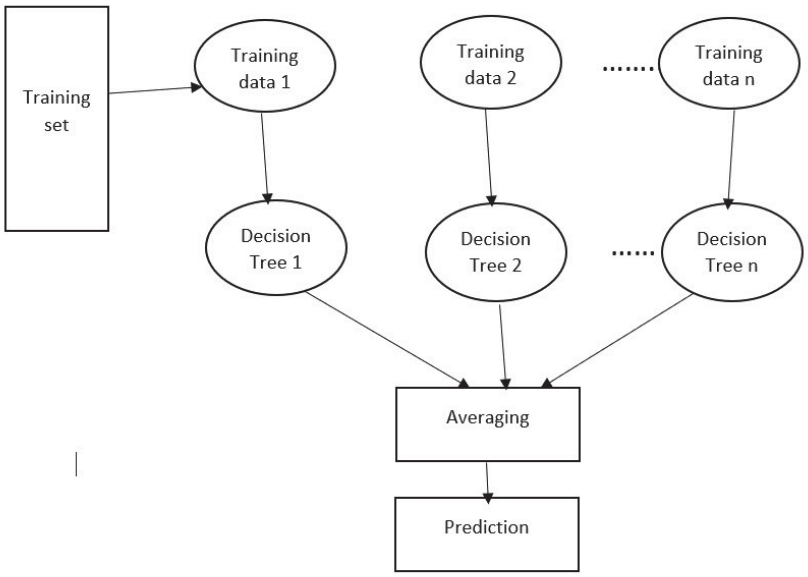


Figure 1.1: Random Forest Classifier

Gradient boosting algorithm where a successive decision tree is formed from the previous decision tree and taking its flaws into consideration. The following figure shows how the gradient booster classifier predicts the results.

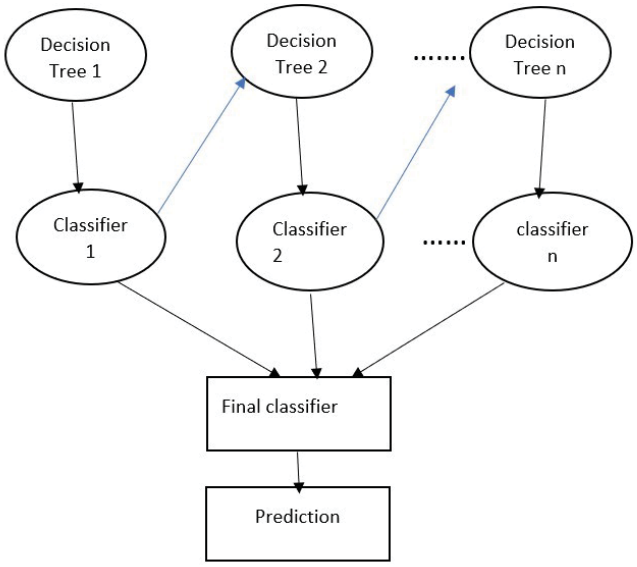


Figure 1.2: Gradient Boosting Classifier

Chapter 2

Related Work

Shlok Gilda in 2017 [7] have performed research on evaluating machine learning algorithms for detection of fake news on media. The data set is obtained from single media and list of sources from open sources.co. This paper mainly explains how language processing technique is used for the detection of fake news. The method used in this research is frequency-inverse document frequency (TF-IDF) detection of about 11,000 articles. The data set is tested on various classification algorithms like Gradient Boosting, Bounded Decision Trees and Random Forests. The results of this research shows that the fake news sources are identified with an accuracy of 77.2 percent for Random Forest algorithm.

Smitha and Bharath in 2020 [19] have performed research on performance comparison of machine learning classifiers for fake news detection on media. The fundamental target of this research is to propose a model that detects fake news by using three types of Natural Language Processing text vectorization and applying on different ML classifiers thus producing good results for classifying news as fake or real. The methodology of this research involves text collection in the first stage. The text is collected from one of the Kaggle data set which consists of around 13000 posts. Then in the later stage text pre-processing, feature extraction is done. The results of this research shows that SVM linear algorithm has the better accuracy of 0.94 in detecting fake news when compared to Decision Tree and Logistic Regression Algorithms.

Hugo Queiroz Abonizio, Janaina Ignacio de Morais in 2020 [1] have performed research on language independent fake news articles detection using machine learning algorithms. In this paper the research is done by comparing four machine learning algorithms namely k-Nearest Neighbors (k-NN), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB). The method of this research involves text processing pipeline, news collection process and classification algorithms. The results of this paper shows that detection of fake news irrespective of language with Random Forest algorithm is comparatively high with an accuracy of 85.3 percent.

Rahul, Mamatha, Shiva, Monica, Krishna in 2020 [12] have performed their research on identification of fake news articles using machine learning algorithms. In this paper the research is done using Machine Learning algorithms such as Naive Bayes, Passive Aggressive Classifier and Deep Neural Networks. These algorithms are used on eight different data sets acquired from various sources. In this paper initially the data is collected and then the data is pre-processed. Then the data sets are then split into train sets and finally tested on test set. Here TF-IDF vectorization is done. The results of this paper shows that Navies Bayes algorithms has the highest accuracy in detecting fake news.

In this paper a comparison is made between Random Forest algorithm and Gradient Boosting algorithm to detect fake news articles on media. Our paper considers data set from Kaggle and Scopus, and it uses Supervised Machine Learning algorithms.

Chapter 3

Method

This chapter gives the overview of all the experimental work done to attain the aims and objectives of this thesis and to answer the research questions. In our thesis comparison of two machine learning classifiers is done on Kaggle data set and Scopus data set.We have taken the Experimentation research method to perform our thesis. The experiment checks which machine learning classifier; either Gradient Boosting classifier or Random Forest classifier is efficient in detecting fake news articles on media. The two specified machine learning classifiers are compared based on their precision, recall, accuracy and F1 score [8]. The complete experimental procedure is as follows:

- In the initial step data collection is done from different sources, here from Scopus and Kaggle.
- In the latter step data is trained and tested.
- TF-IDF vectorization model is used to train the classifier.
- Now, algorithms of the classifiers are applied.
- Selecting the best algorithm based on the obtained results.

3.1 Evaluation Metrics:

True Positive (TP): The total number of positive predicted values and the expected values are true.
True Negative (TN): The total number of negative predicted values and the expected values are true.
False Positive (FP): This is the total number of positive predicted values but the real values are false.
False Negative (FN): This is the total number of negative predicted values and the actual values are also false.

Recall Recall is how many true positives are found. Recall can be calculated as follows: Recall = TP/(TP+FN)

Precision Precision can be defined as the total number of correct positive predictions made. Precision can be calculated as, precision = TP/(TP+FP)

Accuracy This is defined as the percentage of correct predictions with the over all predictions. The classifier with high accuracy is preferred. Accuracy = TP+TN/(TP+FP+TN+FN)

F1 Score It is a way to combine both precision and recall to balance both the properties. F score can be calculated as follows:
F1 score = (2*Recall*Precision)/(Recall+Precision)

3.2 Experimental Design

In this paper Python programming is used for building models for machine learning algorithms. The following are the libraries used to manipulate data in the thesis. To carry out this research it is required to collect the data and train and test the model on that specific data. The following are the steps involved in our thesis model.

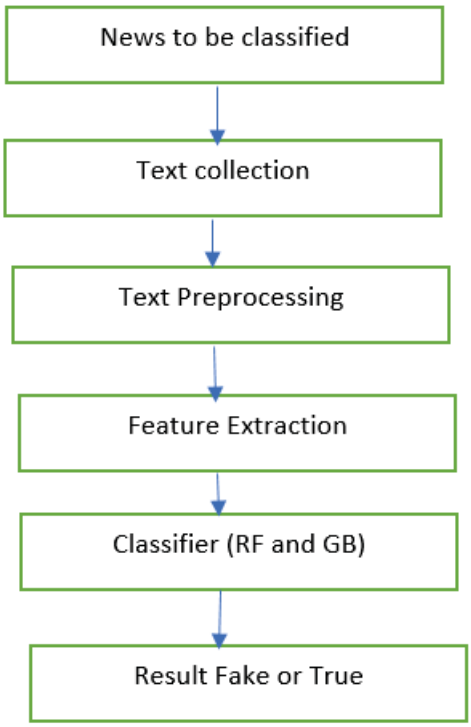


Figure 3.1: Work flow

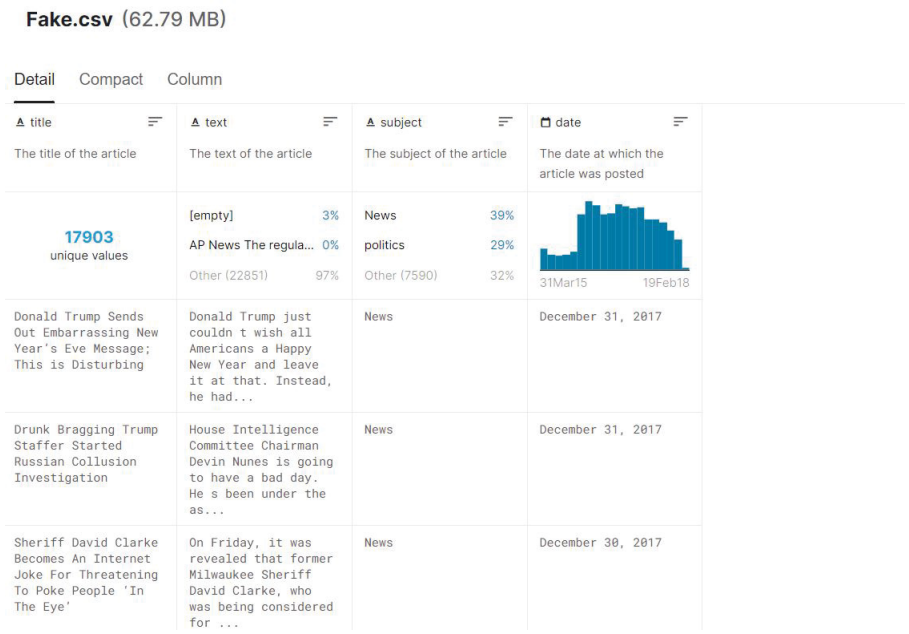
3.3 Data Set Explanation and Processing

Work flow for this thesis the data sets are collected from Kaggle web- site and Scopus. The data sets which we have considered consists of 20000 fake and real news articles.

We process the data according with the following steps:

- All the characters of the news article are converted into lower case letters
- All the numerical values are removed
- Any punctuation marks or accent marks are all removed
- All the blank spaces are removed.
- Repeated words are expelled.

The data set taken from Kaggle data base has combination of both real and fake news articles. The real news articles are in real.csv and fake news articles are in fake.csv and the figure 3.2 and 3.3 below shows the data set over view.

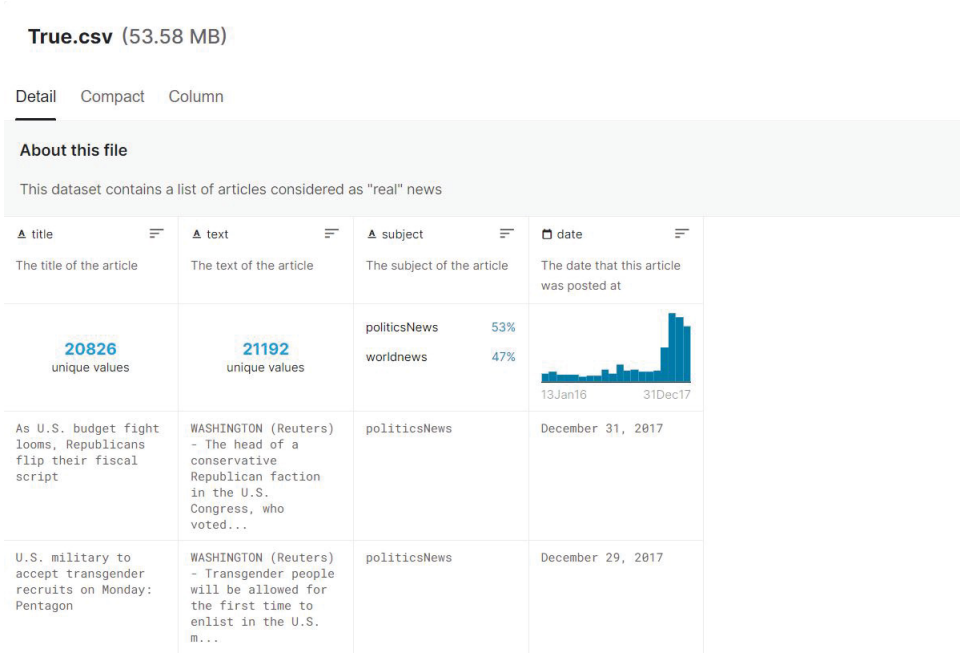


Fake.csv (62.79 MB)

Detail Compact Column

title The title of the article	text The text of the article	subject The subject of the article	date 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
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the as...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People 'In The Eye'	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for ...	News	December 30, 2017

Figure 3.2: Code



True.csv (53.58 MB)

Detail Compact Column

About this file

This dataset contains a list of articles considered as "real" news

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
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. m...	politicsNews	December 29, 2017

Figure 3.3: Code

Training and Testing: The experiment is conducted by training 80 percent of the data set using the classification algorithms random forest and gradient boosting classifier and comparing the results after training the model. The following is the code used for detecting fake news articles.

Feature Extraction: The data needs to be parsed to evaluate words and words should be encoded as integers or floating-point values before giving it into the machine learning algorithm. In this paper we used TF-IDF vectorization model. We are going to create a function to process the text and convert text to vectors using the TF-IDF vectorizer which converts words into meaningful representation of numbers, it assigns a metric to show that the particular word is present. If a word appears frequently in a document, it should be given a high priority. If a word is present in too many other documents, it is not a unique identifier and it should be given a low priority. We standardize the occurrence of a word with the size of the document hence we use term frequency (TF).

To minimize the effect of a word that has high frequency in many documents which contributes less in deriving the meaning of the document we use the Inverse document frequency (IDF). Using this TF-IDF vectorization and machine learning algorithms we can detect fake news articles. To justify the results that have been obtained after training manual testing is performed.

- Input: News content, then convert text to lowercase.

```
#gradient boosting classifier
from sklearn.ensemble import GradientBoostingClassifier
#random forest classifier
from sklearn.ensemble import RandomForestClassifier
GBC = GradientBoostingClassifier(random_state=0)
RFC = RandomForestClassifier(random_state=0)

#fitting the training data to the classifiers
GBC.fit(xv_train, y_train)
RFC.fit(xv_train, y_train)

pred_gbc = GBC.predict(xv_test)
pred_rfc = RFC.predict(xv_test)


print("Gradient boosting Classifier\n",classification_report(y_test, pred_gbc))
GBC.score(xv_test, y_test)

print("Random Forest Classifier\n",classification_report(y_test, pred_rfc))
RFC.score(xv_test, y_test)

if GBC.score(xv_test, y_test)> RFC.score(xv_test, y_test):
    print("Gradient boosting Classifier has the highest accuracy of ",GBC.score(xv_test, y_test) )
else:
    print("Random Forest Classifier has the highest accuracy of ",RFC.score(xv_test, y_test) )
```

Figure 3.4: Code

- Remove punctuation's, digits, stop words from text.
- Repeat, Input: Receive each news article, calculate TF-IDF vector for it. Append the TF-IDF vector to TF-IDF feature vector until end of news articles.
- Parse TF-IDF feature vector into classifier and return feature vector giving the highest accuracy.
- Build model with the feature vector.
- Output: Prediction of news - Fake or Real

Chapter 4

Results

This chapter deals shares the results which we have obtained in comparing the Random Forest classifier and Gradient Boosting classifier to detect fake news articles on media from the datasets taken from Scopus and Kaggle.

These are the data sets we have considered:
https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
https://data.mendeley.com/datasets/945z9xkc8d/1

The features of the taken data set are as follows Features of Data Set1:

- Title
- Text
- Subject
- Date

Features of Data Set2:

- Text
- Label

The following table shows the performance evaluation of the two specified algorithms with different evaluation metrics using TF-IDF vectorization model. The figure 4.1 shows the evaluation metrics Accuracy, Precision, Recall, F1 Score for both the algorithms Random Forest and Gradient Boosting on Data Set1. We have taken two different data sets and trained them individually. It is clear from the results of data set 1 that the Gradient Boosting has high accuracy than Random Forest in detecting fake news articles on media.

Data set 1	RF	GB
Accuracy	0.99	1
Precision	0.89	0.9
Recall	0.89	0.9
F1 Score	0.89	0.9

Figure 4.1: Table with evaluation metrics Data Set1

The figure 4.2 shows the evaluation metrics for both the algorithms on Data Set2. It shows the obtained results for Accuracy, Precision, Recall, F1 Score on data set 2 and the results here depicts that the Gradient Boosting algorithm has high accuracy in detecting fake news articles on media.

Data set 2	RF	GB
Accuracy	0.99	1
Precision	0.99	0.99
Recall	0.99	1
F1 Score	0.99	0.99

Figure 4.2: Table with evaluation metrics Data Set2

The figure 4.3 shows the evaluation metrics for both the algorithms on mixed data set (data set 1 and data set 2). It shows the obtained results for Accuracy, Precision, Recall, F1 Score on mixed data set (data set 1 and data set 2) and the results here depicts that the Gradient Boosting algorithm has high accuracy in detecting fake news articles on media.

Data set 1 + Data Set 2	RF	GB
Accuracy	0.98	0.99
Precision	0.88	0.9
Recall	0.89	1
F1 Score	0.89	0.9

Figure 4.3: Table with evaluation Data Set 1 and 2

The figure 4.4 shows the representation of the evaluation metrics Accuracy, Precision, F1 Score and Recall for both Random Forest and Gradient Boosting algorithms on Data Set 1.

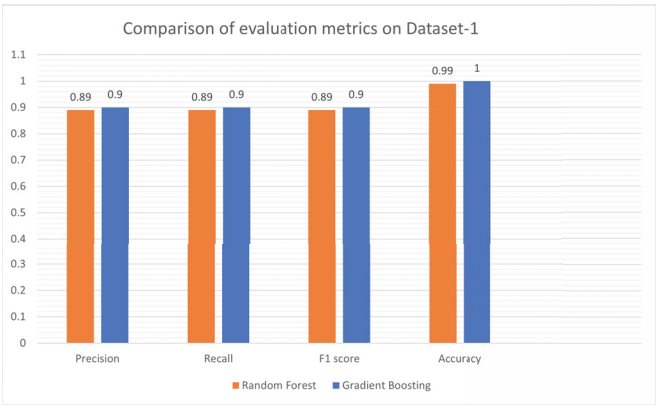


Figure 4.4: Bar Representation of Data Set1

The figure 4.5 shows the representation of the evaluation metrics Accuracy, Precision, F1 Score and Recall for both Random Forest and Gradient Boosting algorithms on Data Set 2.

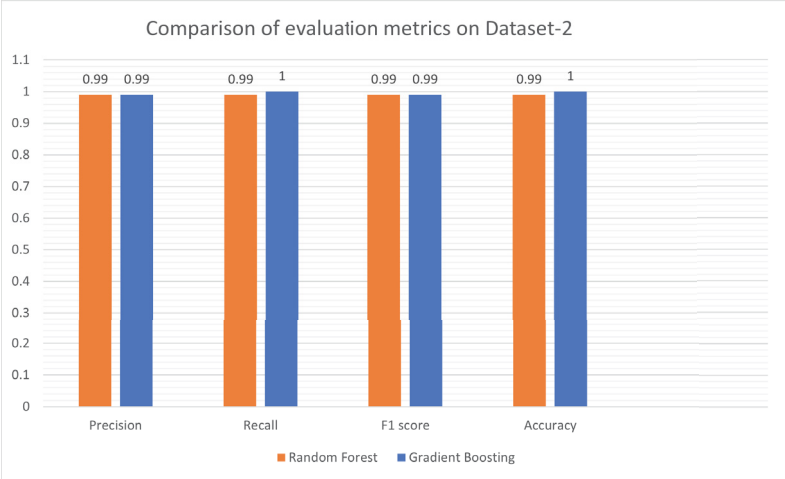


Figure 4.5: Bar Representation of Data Set2

The figure 4.6 shows the representation of the evaluation metrics Accuracy, Precision, F1 Score and Recall for both Random Forest and Gradient Boosting algorithms on Mixed data set (data set 1 and data set 2).

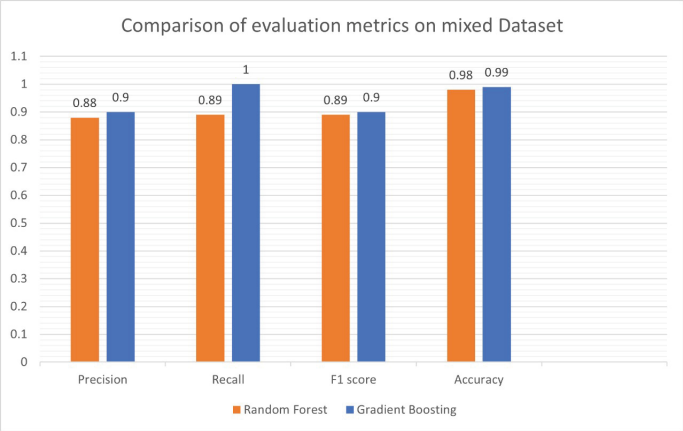


Figure 4.6: Bar Representation of Mixed Data Set

Chapter 5

Discussion

RQ: Which of the machine learning algorithms Random Forest or Gradient Boosting Classifier is efficient in detecting fake news articles in media?

We performed testing on both the algorithms on different data sets and it is the Gradient Boosting classifier which has higher accuracy than Random Forest Classifier in all of the data sets considered. By this observation we conclude RQ as, Gradient Boosting Classifier is efficient in detecting fake news on media.

Machine learning can help to detect fake news on media. Machine learning has different algorithms and various classifiers to detect fake news articles. The classifier with high accuracy will detect the fake news more precisely. The accuracy of the classifier depends as a whole on the training model of the classifier. Training of classifier is the most important task in machine learning. For training the classifiers there are three different models they are, TF-IDF model, N-gram model and Bag of words model. All these models extracts features from the trained data set and then the classifier will go through these features. Thus, machine learning provides better mechanisms to detect fake news with its various classifiers with high accuracy.

Chapter 6

Conclusions and Future Work

With increase in number of fake news articles on media, it is important to detect fake news and in this paper we have compared Random Forest and Gradient Boosting classifiers to detect fake news on media. We have observed that machine learning provides potential to detecting fake news as it has various classifiers to detect fake news. Each classifier has its own accuracy depending on the trained model. we have used TF-IDF vectorization to train the classifier. we have observed that Gradient Booosting algorithm has better accuracy than Random Forest algorithm. Gradient Boosting classifier is the efficient machine learning algorithm for detecting fake news.

For the future work, we would like to compare several machine learning algorithms with various training models. Each classifier has varied accuracy depending on the trained model. we would like to expand our thesis work for larger data sets.

References

[1] Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87, 2020.

[2] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020, 2020.

[3] Gi-Wook Cha, Hyeun-Jun Moon, and Young-Chan Kim. Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *International Journal of Environmental Research and Public Health*, 18(16):8530, 2021.

[4] Issam El Naqa and Martin J Murphy. What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer, 2015.

[5] Pedro Henrique Arruda Faustini and Thiago Ferreira Covoes. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503, 2020.

[6] Robert Gentleman and Vincent J Carey. Unsupervised machine learning. In *Bioconductor case studies*, pages 137–157. Springer, 2008.

[7] Shlok Gilda. Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th student conference on research and development (SCOReD)*, pages 110–115. IEEE, 2017.

[8] Marina Danchovsky Ibrishimova and Kin Fun Li. A machine learning approach to fake news detection using knowledge verification and natural language processing. In *International Conference on Intelligent Networking and Collaborative Systems*, pages 223–234. Springer, 2019.

[9] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[10] Junaed Younus Khan, Md Khondaker, Tawkat Islam, Anindya Iqbal, and Sadia Afroz. A benchmark study on machine learning methods for fake news detection. *arXiv preprint arXiv:1905.04749*, 2019.

[11] Prasad Kulkarni, Suyash Karwande, Rhucha Keskar, Prashant Kale, and Sumitra Iyer. Fake news detection using machine learning. In *ITM Web of Conferences*, volume 40, page 03003. EDP Sciences, 2021.

[12] Rahul R Mandical, N Mamatha, N Shivakumar, R Monica, and AN Krishna. Identification of fake news using machine learning. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE, 2020.

[13] Syed Ishfaq Manzoor, Jimmy Singla, et al. Fake news detection using machine learning approaches: A systematic review. In *2019 3rd international conference on trends in electronics and informatics (ICOEI)*, pages 230–234. IEEE, 2019.

[14] Nicole O'Brien. *Machine learning for detection of fake news*. PhD thesis, Massachusetts Institute of Technology, 2018.

[15] Julio CS Reis, André Correia, Fabricio Murai, Adriano Veloso, and Fabrício Benevenuto. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM conference on web science*, pages 17–26, 2019.

[16] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[17] Jose J Salazar, Lean Garland, Jesus Ochoa, and Michael J Pyrcz. Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy. *Journal of Petroleum Science and Engineering*, 209:109885, 2022.

[18] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

[19] N Smitha and R Bharath. Performance comparison of machine learning classifiers for fake news detection. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 696–700. IEEE, 2020.

[20] HYNDAVI TUMATI, SHALEM RAJU YADALA, NAVYASRI INDUPALLI, and SUNEETHA MANNE. Fake news detection using machine learning techniques.

[21] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.