iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46838

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Fake News Detection Using Deep Learning

Nazakat Farooq Khan[1], Ankur Gupta[2]
[1]*M. Tech Scholar, Department of Computer Science and Engineering, RIMT University, Mandi Gobingarh, Punjab, India*
[2] *Assistant Professor, Department of Computer science and Engineering, RIMT University, Mandi Gobingarh, Punjab, India*

Abstract: Social media news may be a double-edged sword. There are a number of benefits to utilizing it: It's simple to use, takes little time, and is user-friendly. It's also simple to share socially significant data with others. On the other hand, a number of social networking sites adapt the news based on personal opinions and interests. This sort of misinformation is spread over social media with the intent of causing harm to a person, organization, or institution. Because of the prevalence of fake news, computer tools are needed to detect it. Fake news detection aims to aid users in spotting various sorts of fake news. We can tell if the news is genuine or created if we have encountered fake or authentic news before. We may use a number of models to understand social media news. This is a donation in two ways. We must first give datasets containing both fake and accurate news and conduct multiple experiments before developing a false news detector. Various machine learning techniques are used to categorize the data. Random Forest, Logistic Regression, Naives Bayes, Gradient Boost and Decision Tree techniques are used and compared. It was found that Gradient Boost has the best accuracy.
Keywords: Fake news, Deep Learning, Media, NLP

I. INTRODUCTION

Fake news swiftly grew in popularity as a means of disseminating or spreading false information in attempt to influence people's behaviour. The proliferation of false news[2] during the 2016 US presidential elections exposed it as incontrovertible. The following are some facts about false news in the United States. Sixty-two present of Americans get their news from social media. On Facebook, bogus news has a higher share than real news[4]. False news also influenced the "Brexit" referendum in the United Kingdom. In this paper, I investigate the possibility of detecting fake news using traditional learning approaches by just adding text.
Data mining prospects[5] are used to detect fake social media news. The characteristics come first, followed by the measurement. The latter is inaccurate information. In order to construct detection models, characterization must occur before attempting to identify bogus news.
Authenticity and aim are two aspects of the concept of fake news. Authenticity entails the verification of falsifiable information, which implies that the conspiracy theory is not included in the falsified news since it is either false or true in most circumstances. The document's purpose, the second component, consists of writing incorrect facts in order to fool the reader.

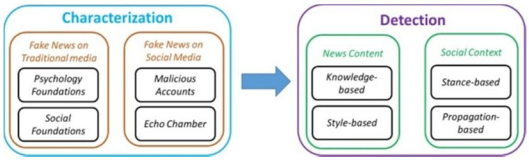


Figure 1. False news on social media: from recognition to detection.[3]

The qualities used to categorize the fake news are four key raw components to consider: They are:
Source: Where the information comes from, who developed it, and whether or not this source can be trusted.
Title: A quick description of the news the reader tries to draw.
Body: The real linguistic substance of the news is written in the body.
Textual content is generally agreed upon alongside visual information, such as photographs, movies, or music.
With verbal and visual core characteristics, these four main components may be reconstructed. As previously said, bogus material is utilized to persuade a customer and is generally written in a way that appeals to the reader. Non-fake warnings, on the other hand, tend to use a more formal language register.
These are linguistic characteristics that can have lexical characteristics due to the total number of words, frequency of words, or specific words. The second consideration is visual elements, aspects of appearance. In fact, manipulated images are frequently utilized to give textual information more weight.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

II. LITERATURE REVIEW

Ruchanskyet al. [11] employed a hybrid low-profile detection technique that included diverse capabilities, such as the temporal interaction of n users with m news items, across time.

Tacchiniet al. [12] has developed a method for detecting false information based on data from social media sites such as likes and users. Thorne advocated a stacked ensemble classification to cope with a false news classification problem. In reality, an article either supports or opposes a fact.

Granik and Mesyura[13] categorize news from buzz data sets using Nave Bayes classifiers. Yang has employed the visual portions of neural networking visuals in addition to text and social characteristics. Wang employs visual cues to identify fake news, but he does it with unfavourable neural networks.

Himank Gupta et. al. [16] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non-spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of-Words model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%.

Marco L. Della Vedova et. al. [17] first proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing its accuracy up to 78.8%. Second, they implemented their method within a Facebook Messenger Chabot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%. Their goal was to classify a news item as reliable or fake; they first described the datasets they used for their test, then presented the content-based approach they implemented and the method they proposed to combine it with a social-based approach available in the literature. The resulting dataset is composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2, 300, 00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes and 6,577 (42.4%) are non-hoaxes.

Mykhailo Granik et. al. in their paper [18] shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may becaused by the skewness of the dataset: only 4.9% of it is fake news.

III. METHODOLOGY

A. Logistic Regression

Logistic regression explains the likelihood of categorization difficulties with two possible outcomes. It is an expansion of the linear regression classification problem model. For regression, the linear regression model works well, but classification does not. Why is this the case? What gives? What gives? One class with 0 for two classes, one class with 1 for one class, and one class using linear regression for one class. Most linear models are weighted, and it works theoretically. However, there are a couple flaws with this strategy: A linear model is unlikely to produce classes; instead, it will treat them as numbers in the ideal hyperplane, minimizing the distance between points and hyperplanes. It just connects items and cannot be interpreted as probability. A linear model also extrapolates and produces values that are below and below zero. There is no significant threshold for differentiating between one class and another since the anticipated result is a linear interpolation of points rather than a probability. Stack overflow is a nice example of this problem. Linear models do not address multi-class classification issues.

1) Advantages and Disadvantages

Many of the advantages and disadvantages of the linear regression model apply to the logistic regression model. Many people have regressed logistically, despite the fact that they are suffering with their restricted expression (e.g. manually formed interactions) and that alternative models can help. Another disadvantage of the logistic regression model is that it is more difficult to understand since weight interpretation is many and does not add up. The logistical regression might lead to total separation. The logistic regression model cannot be trained further if the two groups are fully distinguished. This is due to the fact that the weight for this feature would never converge since its ideal weight was infinite. It’s a shame, because it's such a valuable trait. However, if you have a simple rule that divides both groups, you won't require any machine training. Weight penalization or the generation of a prior probability distribution of weights can be used to solve the entire separation problem.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

On the right, the logistic regression model provides you with not just a classification model, but also an opportunity. This is a significant advantage over models that can only be identified by their finish. Knowing that an instance has a 99 present chance for a class vs. 51 percent makes a major impact.
It may also be converted to a multi-class regression. The Multinomial regression is then triggered.

B. Decision Tree

Linear regression and logistic regression patterns fail when characteristics and outcomes are non-linear or interact with one another. Now is your chance to shine in the decision tree! Data is multiplied by particular cut-offs in functions in tree-based models. A sub-set per instance is used to create different sub-sets. End nodes or feature nodes relate to the last subsets, whereas internal nodes or splits refer to the secondary subsets. To forecast the outcome, the average training results for each node are used. Classification and regression may both be done with trees.

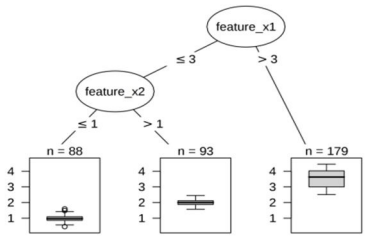


Figure 2:Feature Extraction

1) Advantages

The tree structure enables for the recording of interactions between data components.
Different groups find facts simpler to understand than linearly regressive multi-dimensional hyperplanes. It has an obvious importance, no question. With its nodes and limits, the tree structure provides a natural visualisation.
Because an instance forecast may always be contrasted with the relevant "what if" scenario, a mere node of the tree, the tree explanations are conflicting. The findings are divided into 1 to 3 divisions if the tree is tiny. A three-depth tree just requires three characteristics and split points to represent a specific instance's prediction. The tree predicts the correctness of the forecast. The short trees are relatively simple and general, as each division is easy to grasp with one or two leaves and binary decisions.

C. Random Forest

Random Forest excels at categorization issues [3]. This method was selected for four primary reasons. First, given the numerical and categorical feature set, the notion of traversing a collection of questions using decision trees makes more sense. For example, if the domain score and Facebook popularity indicators are low, it is a solid sign that the news may be untrustworthy. A comparable comparison of the word vector will aid in the identification of a trend in bogus news. Second, the random forest supports a variety of feature types, such as binary, categorical, numerical, and, in particular, the spare matrix, which is utilised to represent the word vector. Third, because random forest employs a collection of decision trees that are trained on a portion of the dataset, overfitting is extremely rare. Overfitting is a challenging problem to detect and correct, and each option to reduce overfitting is a step toward constructing a stronger classifier. Finally, random forest performs well on huge data sets, and as the corpus grows, this is a good approach for the job. It's worth noting that the random forest approach, like any other ensemble algorithm, takes longer to train than popular algorithms like Logistic Regression and Decision Trees. This problem, however, may be solved by employing additional workers in a parallel and distributed system setting.

D. Naives Bayes

It is a powerful classification model that performs well when we have a small dataset and it requires less storage space. It does not produce good results if words are co related between each other [18].

E. Gradient Boosting

The statistical prediction model is another name for the gradient boosting technique. Although it enables the generalisation and optimization of the differential loss functions, it yet behaves relatively similarly to previous boosting techniques. Gradient boosting is typically used in regression and classification processes.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

IV. SIMULATION AND RESULTS

A. *Import the Libraries Necessary for our Project*

```
In [1]: # Fake news Detection

In [49]: import pandas as pd
         import numpy as np
         import seaborn as sn
         import matplotlib.pyplot as plt
         from sklearn.model_selection import train_test_split
         from sklearn.metrics import accuracy_score, confusion_matrix
         from sklearn.metrics import classification_report
         import re
         import string
```

B. *Import the Dataset files and Check Them.*

Inserting fake and real dataset

```
In [2]: df_fake = pd.read_csv("Fake.csv")
        df_true = pd.read_csv("True.csv")

In [3]: df_fake.head(5)
```

Out[3]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn't wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

C. *Change Fake News into 0 and true News into 1.*

```
In [4]: df_true.head(5)
```

Out[4]:

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

Inserting a column called "class" for fake and real news dataset to categories fake and true news.

```
In [5]: df_fake["class"] = 0
        df_true["class"] = 1
```

Removing last 10 rows from both the dataset, for manual testing

```
In [6]: df_fake.shape, df_true.shape
Out[6]: ((23481, 5), (21417, 5))
```

D. *Check the Number of Columns and Rows.*

```
In [7]: df_fake_manual_testing = df_fake.tail(10)
        for i in range(23480,23470,-1):
            df_fake.drop([i], axis = 0, inplace = True)
        df_true_manual_testing = df_true.tail(10)
        for i in range(21416,21406,-1):
            df_true.drop([i], axis = 0, inplace = True)

In [8]: df_fake.shape, df_true.shape
Out[8]: ((23471, 5), (21407, 5))
```

Merging the manual testing dataframe in single dataset and save it in a csv file

```
In [9]: df_fake_manual_testing["class"] = 0
        df_true_manual_testing["class"] = 1
```

E. *Create train and test Dataset*

```
In [10]: df_fake_manual_testing.head(10)
```

Out[10]:

	title	text	subject	date	class
23471	Seven Iranians freed in the prisoner swap have...	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
23472	Whacking Hell & The Fake Left	By Daily Chery and Gilbert Mackler\nWriters ...	Middle-east	January 19, 2016	0
23473	Astroturfing Journalist Reveals Brainwashing ...	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
23474	The New American Century: An Era of Fraud	Paul Craig Roberts In the last years of the 20t...	Middle-east	January 19, 2016	0
23475	Hillary Clinton: 'Israel First' [and no peace ...	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

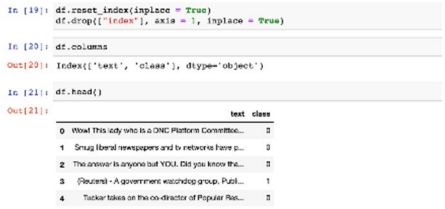
```
In [11]: df_true_manual_testing.head(10)

In [12]: df_manual_testing = pd.concat([df_fake_manual_testing,df_true_manual_testi
         df_manual_testing.to_csv("manual_testing.csv")
```

Merging the main fake and true dataframe

```
In [13]: df_merge = pd.concat([df_fake, df_true], axis =0 )
         df_merge.head(10)
```

F. Add titles to Colomns and Merge Them

G. . Drop the Columns not Needed



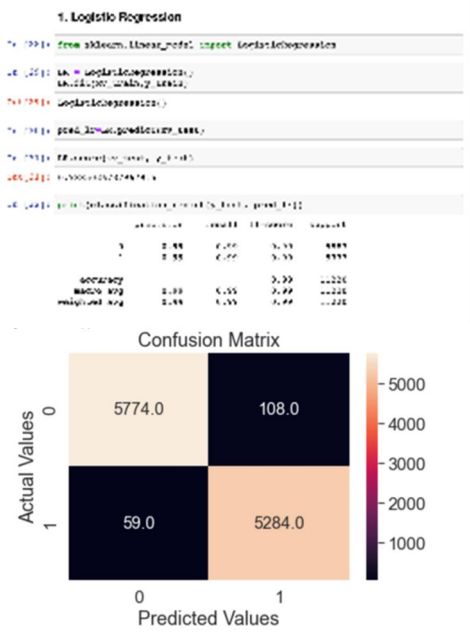
```
In [19]: df.reset_index(inplace = True)
         df.drop(["index"], axis = 1, inplace = True)

In [20]: df.columns
Out[20]: Index(['text', 'class'], dtype='object')

In [21]: df.head()
Out[21]:
```

	text	class
0	Wow! This lady who is a ONC Platform Committee...	0
1	Smug liberal newspapers and tv networks have p...	0
2	The answer is anyone but YOU. Did you know tha...	0
3	(Reuters) - A government watchdog group, Publi...	1
4	Tucker takes on the co-director of Popular Res...	0

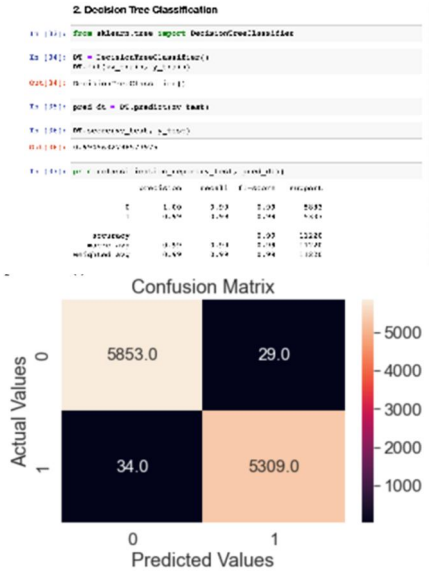
H. Filter the fake and true news by applying Various Processes.

I. Split the dataset , 75% Training and 25% Testing.

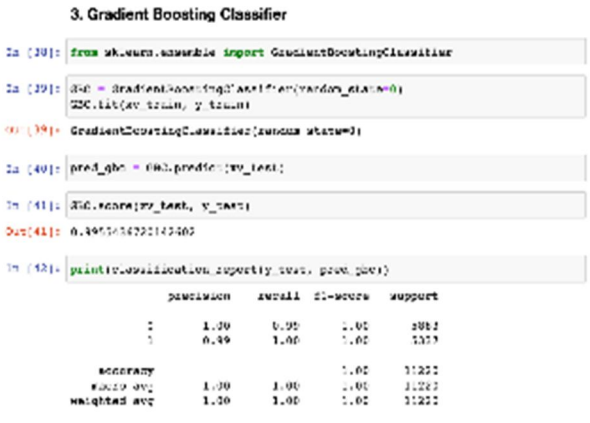
J. Apply Logistic Regression and Check the Results.

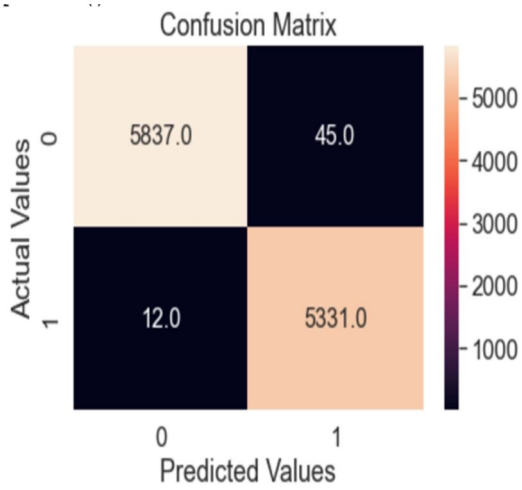


K. Apply Decision Tree and Check the Results.

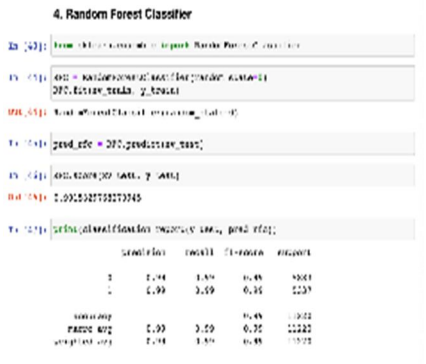


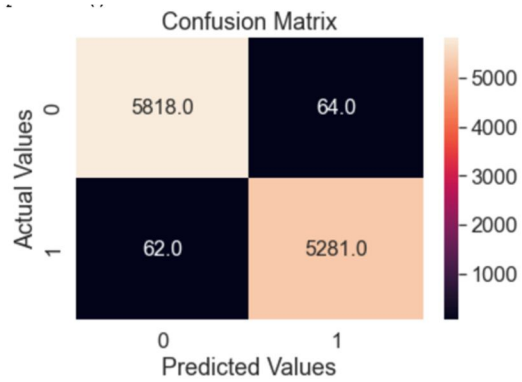
L. Apply Gradient Boosting and Check the Results.





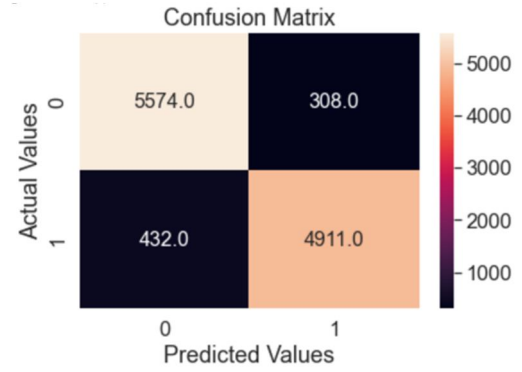
M. Apply Random Forest and Check the Results.



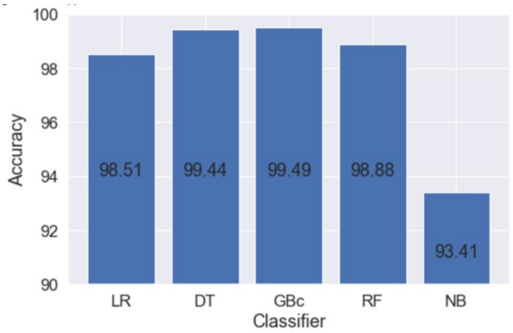


N. Apply Naïve Bayes and Check the Results.

```
print(classification_report(y_test, pred_NB))
              precision    recall  f1-score   support

           0       0.93      0.95      0.94      5882
           1       0.94      0.92      0.93      5343

    accuracy                           0.93     11225
   macro avg       0.93      0.93      0.93     11225
weighted avg       0.93      0.93      0.93     11225
```



After applying all these algorithms, it was found that Gradient Boost algorithm performed the best followed by Decision Tree algorithm.

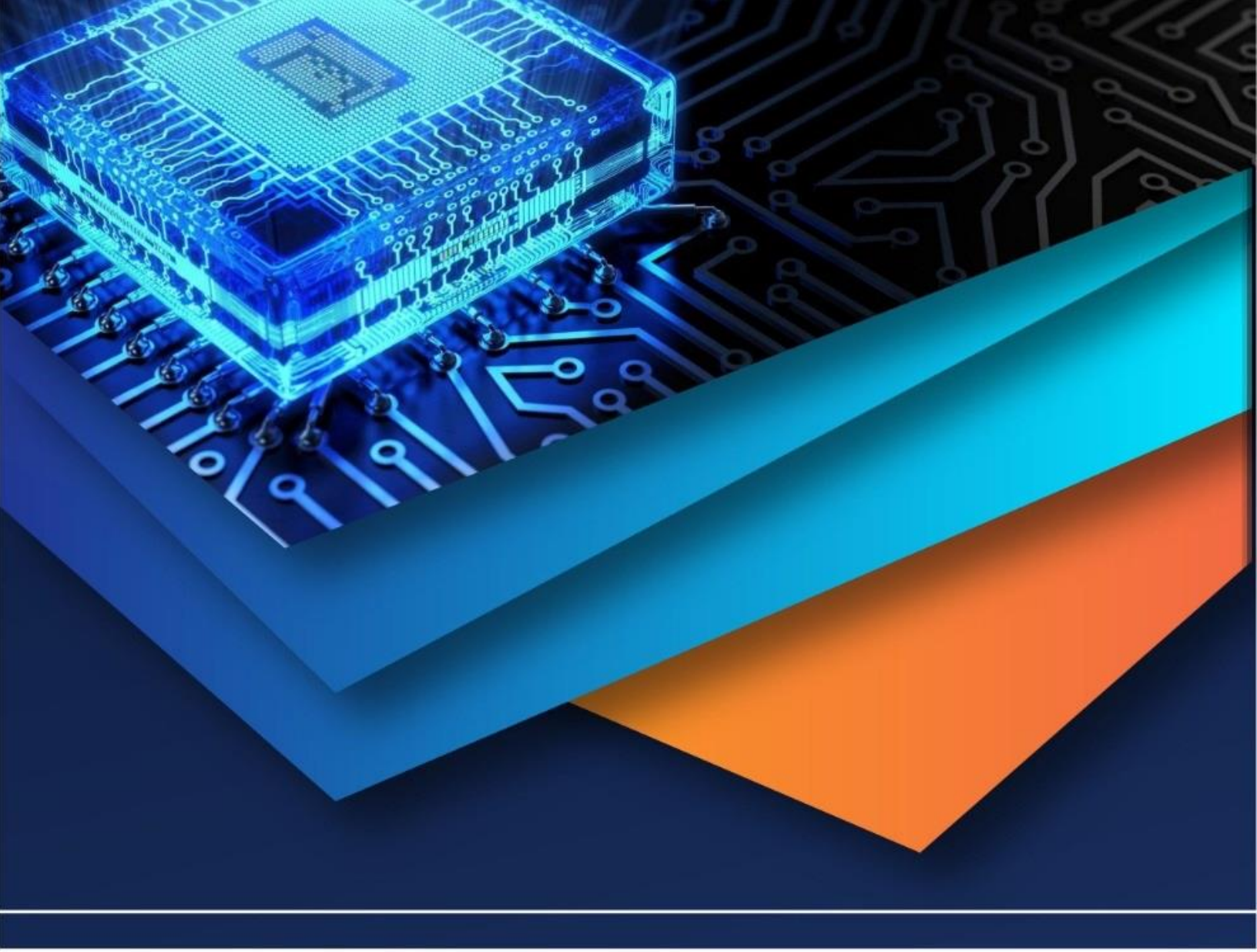


V. CONCLUSION

With more people using the internet, spreading false information is becoming easier. Many people use the Internet and social media on a regular basis. On these sites, there are no limits on posting news. As a result, some people take advantage of these channels and start disseminating false information about people or organizations. This might ruin a person's reputation or have an impact on a business. False news can also influence public opinion about a political party. This fake news must be discovered.
We used five different machine learning algorithms and found Gradient Boost to be the best.

REFERENCES

[1] Great moon hoax. https://en.wikipedia.org/wiki/Great Moon Hoax. [Online; accessed 25-September-2017].
[2] S. Adali, T. Liu, and M. Magdon-Ismail. Optimal link bombs are uncoordinated. In AIRWeb, 2005.
[3] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In ICWSM, 2013.
[4] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017.
[5] E. Arisoy, T. Sainath, B. Kingsbury, and B. Ramabhadran. Deep neural network language models. In WLM, 2012.
[6] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In SIGIR, 1998.
[7] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. In SIGIR, 2007.
[8] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/ libsvm.
[9] K. Chellapilla and D. Chickering. Improving cloaking detection using search query popularity and monetizability. In AIRWeb, 2006.
[10] J. Chung, C¸ . Gulc¸ehre, K. Cho, and Y. Bengio. Empirical evaluation ¨ of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.
[11] E. Convey. Porn sneaks way back on web. The Boston Herald, 1996.
[12] N. Daswani and M. Stoppelman. The anatomy of clickbot.a. In HotBots, 2007.
[13] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In ICASSP, 2013.
[14] Z. Dou, R. Song, X. Yuan, and J. Wen. Are click-through data adequate for learning web search rankings? In CIKM, 2008.
[15] Sharma, Uma & Saran, Sidarth & Patil, Shankar. (2020). Fake News Detection Using Machine Learning Algorithms.
[16] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383
[17] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.
[18] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY