Fake News Detection Using Ensemble Technique in Machine Learning

Yash Mirge
Department of Computer Applications
Veermata Jijabai Technological Institute
Mumbai, India
yashmirge123@gmail.com

***Abstract*—This fake news detection research paper represents use of ensemble learning model for detecting fake news, that means misleading information comes from non-reputable or unwanted resources. Information which is present on Internet, specifically on social media like Facebook, twitter, etc... those are increasingly factors related to fake news. In this research paper we used various ensemble learning techniques for detecting fake news and apply those techniques on online news. These methods use ensemble techniques such as Boosting Technique, Bagging Technique and Stacking Technique to predict whether online news will be fake or real. And we used several ensemble techniques to improve accuracy of our dataset. And this result describes, that is the fake news detection problem related with machine learning methods. In this paper, we demonstrate use of Ensemble Technique. We use fake news datasets, which is textual dataset. Therefore, this study aims at examining the different algorithms and evaluates the efficiency of the algorithms using different performance measures, i.e. accuracy.**

Keywords - Ensemble Learning, Fake News, Bagging Technique, Boosting Technique and Stacking Technique.

I. INTRODUCTION

Fake news is not a new concept. Fake news hide important contextual information. Actually, fake news is a form of news consists of misinformation which spread via media or online social media, etc... [1]. Main interest of researchers are to Classify any news or post like political news, world news, sports news or any other news into fake or real around the world. Stories that looks like a real news but actually are disinformation. Fake news and lack of trust in media are growing problem in our globe. Because many of people focused on classifying online reviews or comments as well as publicly available social media post. Many researchers to find the effect of fake news in our society as well as around the globe. Fake news consists of any type of news i.e. textual or numerical data, news which is actually fake and publisher behaves like a particular news is real and the readers also start will start believing that news but which is not true [2]. For Example, in 2016, fake news spread about American presidential election, at that time people talked about fraudulent voter machine. They took vote using fraudulent voter machine and then they won their election. This news are spread through Facebook, twitter, etc... Basically, fake news are appears on website that looks like a very professional and real, because those news topics or peoples who are trending on Google or Facebook, but sometimes provided information is fake or disinformation.

For fake news analysis, the most commonly used data set has been the Kaggle 5 attribute data set. Most of the publisher refer Kaggle as well as GitHub dataset.

In this work, we aim to present a comparative study of a few existing approaches vs our proposed approaches.

- **How to analyze a news article?**

The below image shows News analysis structure. This fake news information shared by a Facebook user, named of the person was Bob.

Fig. 1. **News analysis structure**

So, this example is categories into four different parts namely as A, B, C and D.

1) A represent as a Creator/Spreader. We can see here in image www.dailynews.com is a source of the news and Bob is anews creator/spreader. Therefore, both we considered as a part of A.
2) B represents as title or heading of the news and another is body of the news which includes as images as well as comments.
3) C represents as reactions of viewer or interaction between users and news articles (means dislikes or likes, comments, time stamps, etc..).
4) D represents any users who involved with this news which can be consider as potential target [3].

Research Problem: Problem statement of research paper is to identify a proper solution which is used to detect

and filter out fake news which will be help to other peoples or users to avoid confusion. So, one of the main challenges which we are facing different types of fake news. Fake news consists of all forms of news which are fake, inaccurate or misleading information which create a big threat in our society [1].

Proposed Solution: The proposed solution is applying machine learning ensemble technique on fake news data set to avoid error rate in our data set, to improve data set accuracy as well as to avoid issues related to fake news.

II. Related Work

Many researchers used several machine learning algorithms like KNN, Naive Bayes, Random Forest, Logistic Regression, Decision tree, Support Vector Machine, etc... to check accuracy of test data set [8]. The author shows that optimization approaches to improve the performance of the algorithm significantly. Random Forest linear model approach was proposed [8]. They aim at creating a classifier that reduces the error rate in the model and then used random forest algorithm [4]. Alternating Decision Trees which combine decision trees and boosting were suggested [8]. XGBoost [8] was used for solving the over fitting problem which usually happens in gradient boosting trees [12]. The author show accuracy of gradient boosting accuracy is 68.7 percent only [8].

Support vector machine is a supervised machine learning algorithm applied on data set for classification as well as regression [3]. Data points are plotted in a multidimensional space [4]. The performance of support vector machine is to find maximum distance between hyper plan and data points to increase accuracy [5]. Advantage of SVM has high performance rate and to create a clear margin between data points. And disadvantages are it take lots of time to train the model as compared to another machine learning algorithm [4]. Therefor accuracy of SVM is 76.2 percent as per author report [8].

Logistic Regression is machine learning regression technique used to make relationship among the variables with the help of statistical methods. This technique is generally used to solved binary classification problem i.e. logistic regression technique deals with predicting probabilities of classes and there are no any restrictions in this technique [4]. This technique also well works on small data set [8].

Random Forest is a machine learning classification algorithm. works on combination of several decision tree and then train the model [4]. The performance of random forest is first to create set of decision trees from subset of the training data set. After that it takes aggregates the results from different decisions trees and then decides final classification of the test data. The author shows that accuracy of these model between 75 to 80 percent only [8].

In Machine learning, for classification most of the research paper used **naive bayes** classification algorithm [1] [5]. Naive bayes classification is based on bayes theorem [4]. It predicts the relationship probabilities for each class such as probability of data points belongs to a particular class [7]. Naive bayes classifier assumes that all features are unrelated to each other [5]. So, the author shows the accuracy of this algorithm on test data set is 68 percent only [8].

So, the reason of choosing this research paper topic is to improve the accuracy of machine learning model with the help of different ensemble learning technique i.e. Bagging, Boosting and Stacking Techniques.

III. Dataset Description

we used data set in our research paper to test accuracy of the model and this data set is provided by Kaggle. Kaggle is most popular machine learning competition platforms. Data set contains news articles or posts labeled as a fake or real news and it has 7000 rows and 5 columns namely as ID, Title, Author, Text and Label.

Data set consists of several type of news like, world news, business related news, economy related news, science and technology, sports, entertainment as well as health. The data set authenticity is checked by Kaggle competition organizers and then they labeled as particular news are Fake or Real.

Title represents brief information in a short way i.e. heading of the newspaper. Text represent detailed information/description of the news like location, people involved in the news, etc... Label represent whether particular news are Real or Fake [6].

TABLE I
Dataset Attributes

Attribute	Representation	Description
News ID	ID	News unique id
News Title	Title	News Heading
News Author	Author	Author/publisher of the news
News Text	Text	Text indicates all fake and real news
Label	Label	1 = Real, 0 = Fake

IV. Methodology

Reason for using python programming language for this research paper implementation because of its simplicity as well as python contains multiple functionalities with various libraries like sklearn, pandas, NumPy, matplotlib, mlxtend, etc...

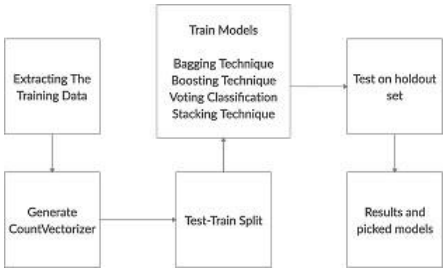

Fig. 2. **System Architecture**

First, we extract training data, after that we will apply Count Vectorizer method and split in train and test data set. Now we train the model using bagging, boosting and stacking technique and based on that we predict accuracy of test data

set. And finally, we picked up particular model which gives better accuracy for our fake news data set.

The Fake News data set is the data set, we use this dataset as the source of weights for our transfer learning model. We use an ensemble model architecture to train the model [10].

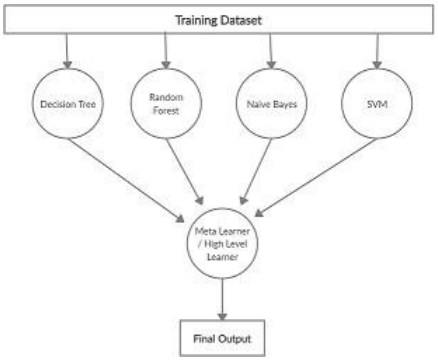

Fig. 3. **Ensemble Model Architecture**

Aim of ensemble learning model is to improve performance of model. The ensemble method is combination of using multiple machine learning model instead of using single machine learning algorithm. We used some real valued function for framework simplicity, which is shown below:

$$g : \mathbb{R}^d \rightarrow \mathbb{R} \tag{1}$$

Ensemble learning model is a process using multiple machine learning model to solve a particular problem. Ensemble is art of combining individual model together to improve stability and efficiency as well as accuracy of the model. There are three main reasons to use ensemble learning model:

1) It provides better accuracy.
2) It gives higher consistency which means to avoid over fitting problem.
3) As well as reduces bias and variance errors.

Single model overfit like a decision tree which is generally overfitted, so in that case we can use random forest algorithm or we can use ensemble of multiple similar models to give us a better fit. The difference in accuracy between the ensemble learning model and individual model is worth the extra training. Then, we can go and use ensemble learning model. And it is possible to use ensemble model for classification as well as regression [9].

Ensemble Technique is categorized into three technique:

Bagging Technique, Boosting Technique and Stacking Technique

1) Bagging Technique:

Bagging is a **parallel learning technique**. Bagging is a part of ensemble learning model where we used multiple models of same machine learning algorithm but with different sub-data set of data picked randomly.

Below mathematical expression shows, train M different trees on different sub dataset and then compute the ensemble:

$$(f(x)) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) \tag{2}$$

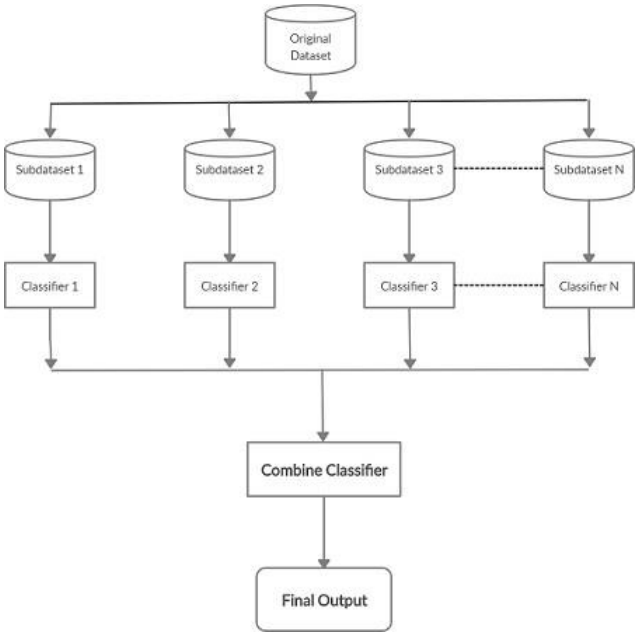

Fig. 4. **Bagging Conceptual Diagram**

Let me make this clearer for you, suppose we have a data set which split into train and test data set for training and validation purpose, like we generally do. In case of Bagging Technique, we select sub-data set of training data set into bags. So, we select those data points randomly, one point at a time and once a point is selected from training data set, then it is not removed from the training data set.

Now, it is also ready to be select data point from training data set. Now, we fill this bag with the sub-data set of training data set and train the model and then we take the vote on their output. This is how exactly bagging works [11].

2) Boosting Technique:

Boosting is a **sequential learning technique.** Boosting is a process where we used machine learning algorithm to combine weak learner to form strong learner in order to increase the model accuracy. Boosting is a little variation on bagging technique. In the case of boosting technique, we select those data points which gives wrong predictions.

We can use this boosting technique for regression as well as classification problems. Boosting technique use on model in a sequential manner.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3}$$

At every single stage we choose decision tree hm(x) is to minimize and a loss function L given the current model Fm-1(x):

$$F_m(x)=F_{m-1}(x)+argmin_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i)+h(x_i)) \quad (4)$$

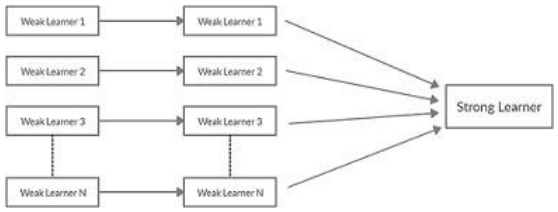

Fig. 5. **Boosting Conceptual Diagram**

In order to improve the accuracy, we select the bag of sub-data set which we are already done for bagging and train the model. we then test the train model with the training data set and picked up each data point which gives wrong prediction. And then we put those data point into second bag.

Along with some randomly chosen data points from the training data set.

Now, we again train the model with the help of new bag of data set and combine it with previously train the model to form an ensemble.

We again test the ensemble of two models on the training data set and again select the points which gives wrong prediction. Along with some randomly selected data points from the training data set and repeat the process until the algorithm can correctly classify the output.

This is how exactly boosting algorithm works [12].

3) Stacking classification Technique:

Stacking is a **Meta Modeling or Split learning Technique.** Stacking contains two type of learner namely as **Base Learner and Meta Learner**. Base learner and Meta Learner are the normal machine learning algorithm like Decision tree, Support Vector Machine (SVM), Random Forest, Naive Bayes, etc...

Base Learner try to fit the normal data set where Meta Learner fir on the prediction of the base learner.

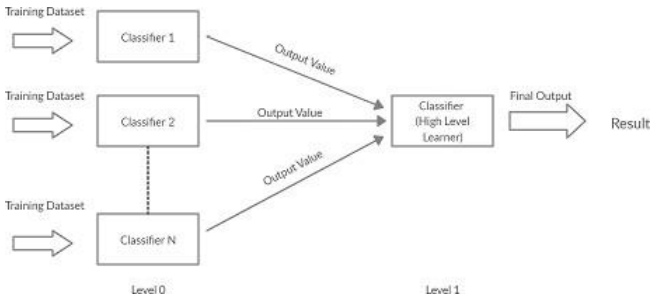

Fig. 6. **Stacking Conceptual Diagram**

I apply multiple machine learning algorithm namely as, K-NN, Naive Bayes, Random Forest, Decision tree and SVM,

TABLE II
ALGORITHM ACCURACY

Algorithms	Accuracy
Bagging	0.901631
Boosting	0.840610
Stacking	0.860321
Naive bayes	0.897417
SVM	0.884354
Random Forest	0.840254
Decision Tree	0.834823
KNN	0.795421

etc... on our data set and predict output. After that we take predicted output and then apply meta-learner or meta-classifier Logistic Regression and then predict final output. Stacking is generally used for winning the Kaggle data science competition.

This is how exactly stacking algorithm works [11].

V. RESULTS

I obtain good results on the Fake News Data set. I apply various ensemble learning technique on that data set. And i really gives a good result.

Fig. 7. Test vs Train Accuracy Using Bagging Algorithm on Fake News Dataset

Fig. 8. Test vs Train Accuracy Using Boosting Algorithm on Fake News Dataset

Fig. 9. Test vs Train Accuracy Using Stacking Algorithm on Fake News Dataset

We can see here, In Bagging ensemble Model classification mis-classification error on test data set starts from 0.175

and ends with 0.100 similarly In Boosting algorithm mis-classification error starts from 0.28 and ends with 0.15 as well as In Stacking algorithm mis-classification error starts from 0.22 and ends with 0.16.

So, we clearly say that bagging gives better accuracy as compared to boosting and stacking technique.

VI. Conclusion

Using Bagging, Boosting and Stacking technique have the following benefits.

1) Reduce the bias in the classification model.
2) Reduce the variance in the classification model.
3) Combining weak learner can result in a strong learner.

We can also conclude the following things:

Here, we compare different ensemble machine learning algorithm. And after conclude that which algorithm is provide greater accuracy for fake news dataset. And we saw here, in implementation bagging gives best accuracy as compare to boosting and stacking as well as other classification algorithm.

Acknowledgment

I would like to show gratitude to the MCA Department, Prof. Kaustubh Kulkarni, whose valuable guidance has been the success of this research paper. Also his suggestions and his instructions has served as the major contribution towards the completion of this research paper. I am also thanks to the institution for providing me with the necessary infrastructure and support to conduct the research on the premise.

References

[1] Monther Aldwairi, Ali Alwahedi. ”Detecting Fake News in Social Media Networks” , Procedia Computer Science, 2018.

[2] Syed Ishfaq Manzoor, Jimmy Singla, Nikita. ”Fake News Detection Using Machine Learning approaches: A systematic Review” , 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019.

[3] Xichen Zhang, Ali A. Ghorbani. ”An overview of online fake news: Characterization, detection, and discussion” , Information Processing Management, 2020.

[4] Ritika Nair, Shubham Rastogi, Tridiv Nandi. ”Fake News Detection”.

[5] Akshay Jain, Amey Kasbe. ”Fake News Detection” , 2018 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), 2018.

[6] Dataset : https://www.kaggle.com/c/fake-news/data

[7] Stefan Helmstetter, Heiko Paulheim. ”Weakly Supervised Learning for Fake News Detection on Twitter”, 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[8] Shlok Gilda. ”Evaluating Machine Learning Algorithms for Fake News Detection” , 2017 IEEE 15th Student Conference on Research and Development (SCOReD).

[9] Akshma chadha, Baijnath Kaushik. ”A survey on prediction of suicidal Ideation Using Machine and Ensemble Learning” , The Computer Journal, 2019.

[10] Shaohua Wan, Hua Yang. ”Comparison among Methods of Ensemble Learning” , 2013 International Symposium on Biometrics and Security Technologies.

[11] Bernard ienko, LjupCo Todorovski, and Sago Dieroski. ”A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods”.

[12] Chien-Hsing Chen, Chung-Chian Hsu . ”Boosted Voting Scheme on Classification ” , 2018.