Ensemble based Approach for Fake News Detection

Parth Patel[1], Shubham Patel[2], Yashkumar Patel[3]

[1, 2, 3]U.G. Students, Smt. Kundanben Dinsha Patel Department of Information Technology, CSPIT, CHARUSAT, Changa, Gujarat

---***---

Abstract - *Fake news detection is a stimulating topic for computer scientists and science. Online users are creating and sharing information more than ever before. There are many fake articles from disparate sources among various users around the world. Most of these articles are misleading, false or have no relevance to reality. It is challenging to automatically classify text articles as rumor or fake. Furthermore, analyzing the text and categorizing the news headlines or false articles for reporting purposes is also daunting. The truthfulness of the text depends on several factors. The work will consist of using an ensemble approach for machine-driven categorization of news articles. Textual properties will be identified using NLTK to summarize the contents of large articles to derive knowledge. The developed ensemble methods based approach will be evaluated with real world data and will provide considerable accuracy and also help to increase prediction power of the individual models.*

Key Words: Ensemble, Fake news, Detection, Machine Learning

1. INTRODUCTION

Misleading information has been a rising problem for many decades especially because of increasing use of social media which is the largest source of spreading fake news[10]. In addition, counterfeit news is generally spread to influence specific groups or politicians. Moreover, the internet is majorly used to spread fake news and people who spread that kind news have intended to damage a person or an entity or an agency to gain political or financial benefits[10]. Furthermore, fake rumors spread to create wrong perceptions of information in the thoughts of people[3]. The growth of fabricated details or phony news is definitely not another marvel. It turns out to be clearer during levels with high media inclusion, like the demonetization measure in India, 2016[11]. Exploration has ordinarily uncovered that out of all online media stages, Twitter does well in uncovering improper information as a result of oneself altering properties of publicly supporting as clients share notions, theories, and proof[11]. It has been observed that, nowadays, numerous types of fake news are currently present on different types of social media objectives and it is a very hard task to tackle them[3]. There are countless systems which can detect whether the news is real or fake, are available but every system has content related issues. The first priority to classify the dataset is the efficient detection systems. In this paper, we assess the exhibition of ensemble approach for counterfeit news recognition on two datasets, one containing customary online news articles and the second one, news from different sources.

2. METHODOLOGY

2.1 PREPROCESSING

2.1.2 Normalization: It is a process of cleaning text by lowercasing to reduce size of the vocabulary of our data, removing or replacing numbers, removing punctuations, unnecessary white spaces and default stopwords[6]. These components have no influence on the interpretation of a sentence. For each language, the NLTK library provides a bunch of stopwords that can be utilized to eliminate them from text and return a rundown of word tokens[1].

2.1.2 Tokenization: It is a process that splits an input sequence into individual meaningful chunks called tokens. These tokens are useful units for further semantic processing[1]. It can be a word, sentence or paragraph, etc. There are multiple tokenizers in NLTK libraries such as WhiteSpaceTokenizer, WordPunctTokenizer, TreebankWordTokenizer, etc. WhiteSpaceTokenizer splits the input using white spaces and WordPunctTokenizer uses punctuations to separate words whereas TreebankTokenizer follows different grammar rules to tokenize inputs[6].

Input = [“ensemble based approach for detection of fake news using machine learning”]

Output = [“ensemble”, ”based”, ”approach”, ”for”, ”detection”, ”of”, ”fake”, ”news”, ”using”, ”machine”, ”learning”]

2.1.3 Stemming: It is a procedure to get the root form of word, in which suffixes are removed or replaced, that is called the Stem[1]. It collects a bunch of words to the same block (stem). There are various stemmer available in NLTK library, namely PorterStemmer, LancasterStemmer, etc. PorterStemmer is a less aggressive algorithm with five rules for different situations, which are gradually applied to build basic knowledge[1,6]. It generates stems that are understandable but often they are not actual English words. Also, rather than keeping a lookup table, it simply applies rules which are based on algorithms, to generate stems. . LancasterStemmer is a type of algorithm which has iterative behavior[6]. Rules are saved externally in LancasterStemmer. There are chances of over-stemming due to more number of iterations, which can either make stems non linguistic or they might not have meaning. Apart from that, it can make it more confusing to understand.

2.1.4 Lemmatization: By using vocabulary and morphological analysis, we usually make things proper[1,6]. That is, the base or dictionary form of a word is returned by this algorithm, which is called lemma. NLTK provides WordNetLemmatizer which has access to a WordNet database to search for lemmas of particular words. For example, a lemmatizer maps gone, going and went to its canonical form go.

2.2 ENSEMBLE LEARNERS

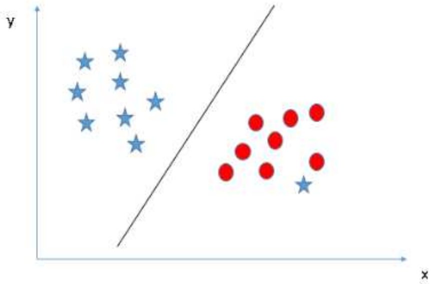
2.2.1 Naïve Bayes: It is one of the machine learning algorithms used for text classification problems[12]. Apart from that, it is very easy to implement and very efficient at the same time. There are three event models:

- Multivariate Bernoulli Event Model
- Multivariate Event Model
- Gaussian Naïve Bayes classification

In addition to this, Naïve Bays means that all features are independent from each other and that occurrence of one feature does not affect the probability of occurrence of another feature. Furthermore, this model outperforms all the powerful models, when it is assigned to do tasks with the small dataset[12].

In multinomial naïve bayes, there is a feature vector which is having a term and that term represents the occurrence of the given term that is frequency. On the other hand, Bernoulli is a binary classification which tells us that a term is present or not and the Gaussian classifier is for the continuous distribution.

2.2.2 SVM (Support Vector Machine): SVM is one of the supervised machine learning models and is used as a classification for regression[7]. However, it is highly used in classification problems. In this algorithm, we usually have value of each and every data in the form of a point on the N-dimensional space where N is a number of spaces we have[7]. And it also carries the value of each element, which represents the value of a specific coordinate[13]. At the end, we find the hyper-plane to classify the values. For instance, in identifying the hyper-plane, the SVM algorithm has the ability to ignore the outliers[12].



There are many easily understandable advantages of SVM Algorithm:

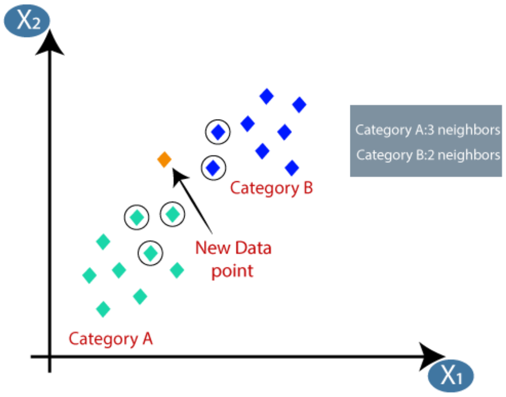
- It works very well when it comes to classification of classes which have the clear and far margin.
- Very effective in large spaces[12].
- Also, efficient if the numbers of dimensions are greater than the numbers of samples.
- SVM is memory efficient.

2.2.3 Logistic Regression: It is a very basic classification model of the Machine learning domain. It is mainly used to predict the binary output such as 0 or 1, Yes or No[7,12]. It takes input of the independent variable. It is also considered as the special case of the linear regression model when the predicted variable is a binary variable. In simple words, the probability of the occurrence of an event is predicted by this model. The probability is always into the range of 0 to 1.

As it is a very basic model it is having very basic advantages:

- It is very easy to implement.
- It is strongly against assumption.
- It can be expanded to multiple classes.
- Very fast when it comes to working with the unknown records compared to others.

2.2.4 KNN (K-Nearest Neighbors): It is one of the only simplest ML models and complies with supervised learning as well. It predicts the similarities between the data for which we are predicting the class and the existing classes and at the end[7]; it put the new case into the category of the class which is very similar to the record or data. It can be used in regression and in classification as well. However, it is mostly used in the classification[12]. Apart from that, It is also a Lazy learner as it is not learning or using a training dataset for memorization , it actually starts performing as soon as it gets data which it needs to predict and that is why it is not using any memory. It is also good at outliers for example:



Here the new data point can be considered as the outlier and can be easily classified into the class A as we are using KNN here. Talking about the advantages:

- First of all it is very easy to implement.
- Data can be added at any time as it is a lazy learner.
- It requires no training period so it is time effective.

2.2.5 Random Forest algorithm: This is a popular supervised machine learning algorithm. For both purposes, whether for classification or for regression, it can be used[7]. It is clearly based in the method of Ensemble learning which is combining the results of multiple classifiers to improve the overall accuracy. Most important aspects about Random Forest algorithm are :

- Time taken by this algorithm is less compared to other algorithms.
- Though we give large dataset to it , it will always give high accuracy among all models.
- It is possible that sometimes data is missing from the data set even though it gives high accuracy.

So, from above statements we can say that it is highly capable to handle the high dimensional dataset and work for both classification and regression. However, it is more suitable for the regression problems.

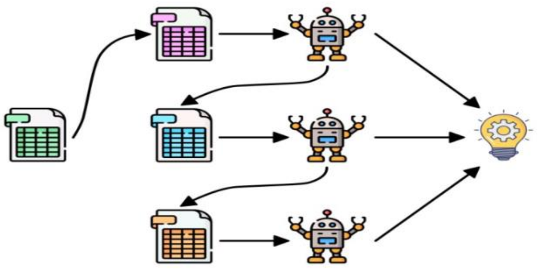
2.3 ENSEMBLE APPROACH

To increase predictive power of first level models, ensemble approach combines all of them. Moreover, ensemble learning approaches are known as second level models.

Certain models perform good in displaying the information, whereas another perform well in demonstrating others[2]. Discover many individual models and combine all the results to arrive at the final solution. The overall robustness of each individual model compensates for the deviations and biases of each model. This provides a composite prediction where the final accuracy is better than the accuracy of individual models. There are two Ensemble methods:

2.3.1 Sequential ensemble methods:

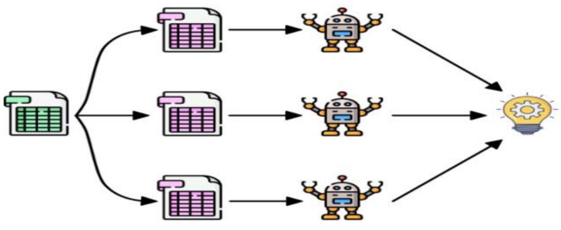
- Individual models come consecutively
- The main motive is to take advantage of the ensemble learners (Individual Models)[2].
- The general overall performance of an ensemble of learners may soar.



Sequential

2.3.2 Parallel ensemble methods:

- All individual models come parallel and execute at the same time.
- Final result considers which has the highest average probability of class.



Parallel

Ensemble model is the application of multiple models to obtain better performances than from a single model.

- Robustness:- Ensemble models incorporate the predictions from all the base learners.
- Accuracy:- Ensemble models deliver accurate predictions and have improved performances[2].

Every individual model has advantages and disadvantages according to the dataset. Any individual model cannot find all the patterns from the dataset so ensemble approach is the best way to combine all the advantages of every individual model to get desired accuracy and to increase predictive power.

Ensemble Learning has multiple methods from which some are mentioned below:

i. Voting: It is a parallel ensemble method.

- Step 1:- Provide original training data.
- Step 2:- Build and fit different classifiers to each of these diverse copies[2].
- Step 3:- To make the final overall forecast, make as many forecasts as possible.

In voting, we give detailed dataset to ensemble learners (Individual models). All ensemble learners are using different models in voting. As it is a parallel method, all ensemble learners provide their output at the same time and the ensemble takes final output which is in majority (individual o/p= 1 1 1 0, Final o/p= 1). There are two types of voting i.e. hard voting and soft voting.

- Hard voting is what I explained above in voting while soft voting has quite similar taste of bagging and boosting.
- Hard voting is totally based on majority while soft voting is based on summing the predicted probabilities of classes and predicting the class with largest sum probabilities.

ii. Bagging: It is a parallel ensemble method. Bagging or bootstrap aggregation[2] reduces variance of an estimate by taking mean of multiple estimates.

- Step 1:- Create randomly sampled datasets of the original training data (bootstrapping).
- Step 2:-Build and fit classifier to each of these diverse copies.

- Step 3:- Take the average of all forecasts to create the final overall forecast.

In bagging, we give subset datasets to our ensemble learners (Individual models) which are randomly picked data. All ensemble learners are using the same model in bagging. As it is a parallel method, all ensemble learners provide their output at the same time and the ensemble takes final output which has the highest probabilities (soft voting).

iii. Boosting: It is a sequential ensemble method and reduces bias by training weak learners sequentially, each trying to correct its predecessor. Boosting is a method for transforming frail learners into solid ones[3]. Each new tree is based on a slightly altered version of the first dataset.

- Step 1:- Train a classifier H1 that best classifies the data with respect to accuracy.
- Step 2:- Identify the regions where H1 produces errors, add weights to them, and produce a H2 classifier.
- Step 3:- Aggregate those samples for which H1 gives a different result from H2 and produces H3 classifier.

Boosting has two methods listed below:

- AdaBoost is the first boosting algorithm to be adapted in solving practices. Misclassified weight records get updated. If models predict not well, weight increases for wrong predictions and vice versa. Learning occurs with the aid of using altering weight. Its tree normally has 2 leaves.
- Gradient boosting is a technique for progressively, additively, and consecutively training several models. It uses a gradient descent technique to reduce a model's loss function (MSE) by adding weak learners. In algorithm, Generate a basic model, an average ensemble learner or the most commonly used class. Analyze the residual error based on the predicted average value and real value. Now, we create an additional RMI model that uses residuals as a target[3]. We have a new prediction residual. Now, we will calculate the new target predicted value. Thereafter, that we have the residuals (actually predicted), the new RM 2 model will match the target residuals another time and forecast the new residuals.

3. PROPOSED SYSTEM

1. Dataset Description: The proposed method was evaluated based on a data set containing headlines from various news sites. The fundamental goal of this dataset was to group the quantity of phony and genuine rumors[3]. Dataset is balanced as it contains all types of instances like political, social, healthcare, immigration news etc. Below is a depiction of the existing features of the news dataset. We have total instances around 23000. Attributes of dataset -

- News_id – the index of individual news.
- News_url – source of distributor of news.
- Title – The short title text should arouse the reader's interest and introduce the main theme of this article.
- Result – label of particular news(real or fake)

2. Feature Selection: We have implemented below methods to convert string to integer because all ensemble learners work with only integers.

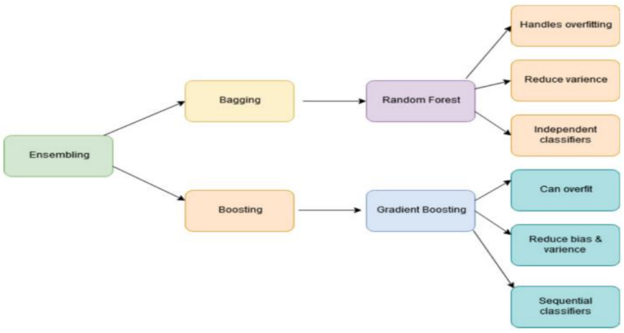
- TF-IDF: Term-frequency times inverse document-frequency. We can say that it defines a calculation of how relevant a word is to a text document[8]. It simply assigns the unique values to features and tells us about the rareness of the word[8].

$$tf - idf(t,d) = tf(t,d) * (\frac{n}{df(t)} + 1)$$

- Count-Vectorizer: It makes a grid where every interesting word is addressed by a column of the words (vertically), and each text sample from the document is a row in the matrix. The value of each cell is nothing but a frequent of particular words in that particular text document[3].

3. Parameter Selection:

- Multinomial naïve-bayes(alpha)
- Logistic Regression (C=100)
- Linear SVM (C=0.25)
- KNN(neighbor=120)
- Random Forest Classifier (number of feature=4)



4. RESULTS

Dry run was performed on several classifiers and their corresponding results were noted[5]. Our observations are appropriately illustrated in this article. Figure 1 display the The classification performance of experiments with different machine learning models (Multinomial Naive Bayes, Logistic Regression, Support Vector Machine, K - Nearest Neighbors, Random Forest)[2,3]. In our evaluation, we tracked down that all machine learning classifiers accomplished approximately 85% and above accuracy.

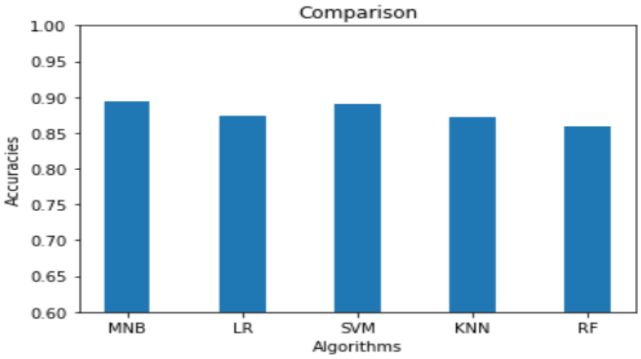


Fig.1. Accuracy

We had implemented svm bagging, Gradientboosting and Hard Voting and got accuracy 89%, 87%, 88% respectively. According to results, bagging with svm took crown place with maximum accuracy among three which means parallel ensemble approach is more beneficial for our dataset. We also measured the confidence score for all and it varies between 0 to 1 because it shows probability. If data with a high confidence score it means that it has high probability to take a place as a final result. Exactness is characterized as the exhibition of our learning calculation for right expectations[5].

Figure 2 and 3 shows the performance of the models. Apart from that it also indicates that the SVM Bagging and the Hard Voting gave output very clear and very effective also, with the TF-IDF[3]. From the figure below, we can get to know the training time of the individual algorithms varies from model to model. We can observe that training time is inversely proportional to number of iterations.

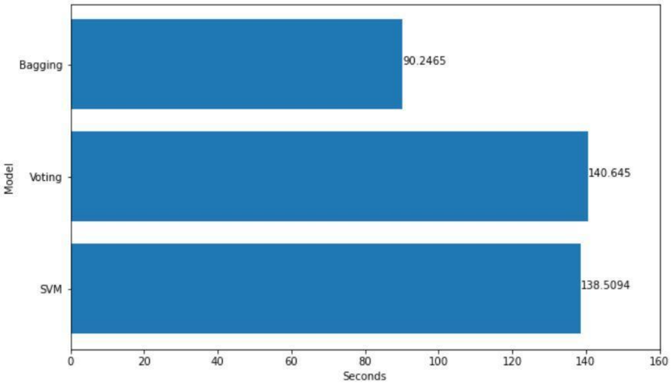


Fig.2. Time taken models

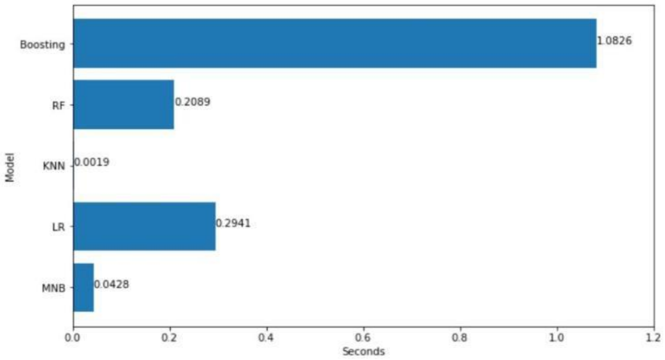


Fig.3. Time taken models

5. CONCLUSION

We are trying to create a machine learning model with the required accuracy to classify real messages and fake messages in our work. It gives an overall methodology and different components over which the believability of information relies on. We implemented monitoring and deep learning methods in our project. The assignment of grouping news physically needs top to bottom information to recognize oddities in the content. In this research, we talked about the complication of counterfeit headlines utilizing ensemble methods. The information we use in our work is collected from the Internet and has headlines in various fields in order to get the most news. An important part of this research is to identify patterns in text data that distinguish fake articles from real news[3,9]. Bagging which is a method of ensemble learning got 89% accuracy to recognize whether news is fake or real, which is 4% higher than the Random Forest Algorithm which is the worst performance among all individuals.. Moreover, it has been extracted from research that ensemble approach performs well rather than individual models.

ACKNOWLEDGEMENT

We have discovered this uncommon chance to manifest an expression of gratitude to every one of the individuals who assumed a key job in the fruitful consummation of our undertaking.

We offer profound thanks to mentor Prof. Hemant Yadav inner task manage from Faculty of Engineering, CHARUSAT for their significant recommendations, help and good help. We likewise thank to every one of the individuals who couldn't locate a different name yet have helped legitimately and in a roundabout way.

We extend our sincere thanks to our external guide Dr. Abdul Jhummarwala (Research Scientist, BISAG-N, Gandhinagar) whose deep knowledge of the subject and helping attitude has provided us with strength to achieve the objective. He was our true guardian during this period with just the right mix of discipline and immense care.

We earnestly thank our Head of Department Dr. Parth Shah for giving the opportunity just as help for constantly being. Furthermore, his capable direction and consistent consolation made us work in all the difficulties during venture advancement.

REFERENCES

[1] C. V. Gonzalez Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 2086-2090, doi: 10.1109/ICDE.2019.00245.

[2] Arush Agarwal, Akhil Dixit. “Fake News Detection: An Ensemble Learning Approach”. Published in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 19 June 2020

[3] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang. “Multiclass Fake News Detection using Ensemble Machine Learning”. Published in 2019 IEEE 9th International Conference on Advanced Computing (IACC). Tiruchirappalli, India. 30 January 2020.

[4] Weijie Jiang, Xingyou Wang, Zhiyong Luo.”Combination of convolutional and recurrent neural network for sentiment analysis of short texts.” COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers 2016.

[5] Ghosal, D., Bhatnagar, S., Akhtar, M.S., Ekbal, A. and Bhattacharyya, P., 2017. IITP at SemEval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval2017) (pp. 899-903)

[6] D. H. Deshmukh, T. Ghorpade and P. Padiya, "Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), 2015, pp. 1-6, doi: 10.1109/ICCICT.2015.7045674.

[7] Ghosh, Souvick, and Chirag Shah. ”Towards automatic fake news classification.” Proceedings of the Association for Information Science and Technology 55, no. 1 (2018): 805-80.

[8] A Aizawa The feature quantity: an information-theoretic perspective of tfidf-like measures Proceedings of the 23rd ACM SIGIR conference on research and development in information retrieval (2000), pp. 104-111

[9] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", Complexity, vol. 2020, Article ID 8885861, 11 pages, 2020.

[10] Figueira Alvaro, Oliveira Luciana.”The current state of fake news: chal- ́ lenges and opportunities.” Procedia Computer Science. 121 (2017),pg 817-825.

[11] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.

[12] Ebtihal A. Hassan, Farid Meziane. "A Survey on Automatic Fake News Identification Techniques for Online and Socially Produced Data" 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2019