ASEAN Engineering Journal

Full Paper

ENSEMBLE CLASSIFIER BASED FAKE NEWS IDENTIFICATION IN ONLINE SOCIAL NETWORK

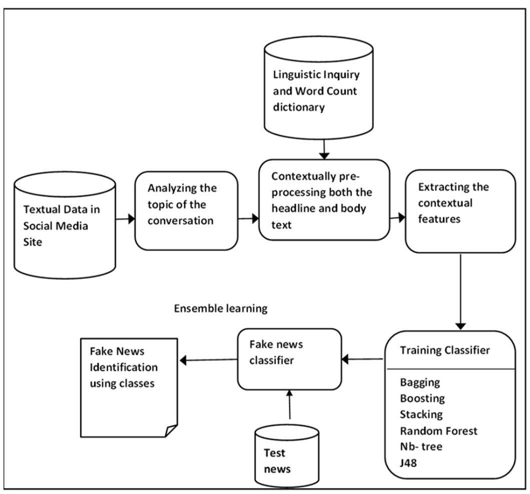
Article history
Received
4 January 2022
Received in revised form
15 March 2022
Accepted
05 July 2022
Published online
29 February 2024

Sandrilla R[a], Savitha Devi[b]

[a]Department of Computer Science, Periyar University, Salem, Tamil Nadu, India
[b]Department of Computer Science, Periyar University College of Arts and Science, Harur, Tamil Nadu, India

*Corresponding author
sandrillashc@gmail.com

Graphical abstract



Abstract

In today's context, one of the main mediums of news consumption is through social media. It has become a common trend now to produce fake news for speedy propagation and popularity. Thus, it creates a sort of illusion and deception for readers. Amidst the umpty number of researches done, we could say that there is no research that could accurately predict fake news over online. In this study, the ensemble classifier is employed to develop a model for fake news identification in online social networks. This proposed method applies the stacked ensemble classification model, the proposed approach that learns the represented text model and classifies the textual data into real news and fake news. Then, the proposed approach takes the collaborative decision from the classified data that is generated from the multiple base learners using the weight-based ensemble method. The accuracy prediction and performance evaluation of time consumption of detection fake news are consecutively 80 Percentage and 11 ms respectively. Thus, it improves the efficiency and classification accuracy over large-scale social media messages through an efficient sentiment analysis model.

Keywords: Stacked Ensemble; NLP; Random Forest; Machine Learning; NBTree; Fake News detection

1.0 INTRODUCTION

The term fake news is defined as the fabricated articles that are intentionally shared on social media to spread false information in order to manipulate reader's opinion [1]. With the increase of internet users nowadays, most of the people consume news through social media. The usage of social media for news updates is a double-edged sword [2]. since it paves a broad road to freely express one's view, on the other hand, it widely opens its door for the dissemination of false news. Fake news is created as extremely significant and has the capability to spread extraordinarily fast through social media. The knowledge of fake information on various topics is rapidly growing and is spreading faster, but overall, the more technical and complex a topic is, the harder it is to produce false claims and information for it [3]. Due to the increase in false news, incorrect information has become a crucial problem in today's society. Social media avails the exposure to new information and stories every day to its increasing numerous users. Thus, Misinformation could be difficult to counterbalance and may have lasting implications [4]. The advent of Web 2.0 technologies and social media enables users to express their own opinions and experiences over Internet. With the volatile growth of a vast amount of textual content in the Web 2.0 applications, news portals, and social networks, there is a huge need of analyzing and extracting knowledge from the posted textual content due to the richness of opinions and attitudes in the user-generated content [5]. The present technologies create easiness for social media users. Hence, the individual creates information, posts, news and promotes faster circulation. This adoption to the internet has been creating less quality with the rapid dissemination of fake news [6]. In this present scenario of online data with excessively with several competing origins of unpredictable quality, it becomes tough for readers to assess the trustworthiness and reliability of what they observe on the internet. [7]. Online social media platforms like Whatsapp, Twitter, and Facebook and have often challenged inspection for being unable to control the spread of false news. Today, social media websites are used as the most fascinated tools for seeing real-time knowledge pertaining to emerging threats, social effects, product styles, and epidemics [8]. Nations or groups have been using the news media to carry out propaganda or influence operations for hundreds, hence fake news itself is not a new problem. The rise of network-generated news on social media makes fake news a more powerful potency group that contests traditional journalistic ethics. In that respect, there are various features of this problem that make it uniquely challenging for automated detection [9]. In fake news detection, determining the truth from the fiction is a quite challenging task, which is in the form of deception detection. Social networks provide the platforms to share information and articles without moderation or fact-checking. Owing to the voluminous and variety of posted information on Social Media Sites, moderating user-generated content becomes an arduous task. Most of the fake news identification approaches transform the identification problems into classification problems using machine learning-based classifiers. Although, there is an essential need to contextually solve disambiguation problems created by the keyword matching methods. Most notably, to discern the veracity of the shared source information, the readers interact with both with the headlines and body of the text. Hence, it is essential to analyze both the headline and the body of the text to determine the fake news. Recently, ensemble learning methods have received greater attention in a variety of applications. Even though ensemble methods improve the classification accuracy in fake news detection, sequential decision making is a fundamental challenge while using the diverse multiple base learners. Also, the traditional sentiment analysis-based classification model lacks accuracy in predicting the intention of the content generated by the social media users due to the features in terms of words in the shared information is not always independent

2.0 RELATED WORK

Fake news detection has become a central research topic in the news industry due to the need of assessing the veracity of digital content over the constant spread of false information [10]. The main aim of the sensationalism of inaccurate eye-catching and intriguing headlines is to retain the attention of audiences throughout all kinds of information broadcast [11]

2.1 Fake News Detection

Chen et al [12], has explained the methods of the rapid spread of fake news using clickbait. The usage of textual and non-textual cues of click-baiting gives the authenticity of the web page regarding the accuracy of the information. In the subject matter and genre regarding the headlines, the structures of grammar and language flow are analyzed using lexical analysis to detect the accuracy of the data [13]. In the attention, certain news sources incorporate technology, for example, Twitter, Facebook, and other user submissions on BBC or ABC blog spots [14], This encourages users to collaborate on on-spot footage and eyewitness interpretations. From some automated approaches, it is identified that list of stylometric features of text (grammatical and language flow) can be utilized to distinguish between two journalistic presentations of test magazines. This distinction was found to be existent regardless of topic, with a predictive accuracy of 77% [15]. Also, the other researchers are analyzing the content based on the connections among news magazines, journalists, and news matters, using deep diffusive network technique has been subjected to integrate information with network structure towards model learning. Zhang et al, [16], have developed a novel diffusive component model, termed as Gated diffusive unit (GDC). This technique consents several inputs from various sources at a time, to efficiently combine this input for the creation of output with news matter as “forget” and “adjust” gates. Mohammad et al, [17] have considered the Stance Dataset tweets interpreted for stance towards trained targets and divergence of language. Thru a stance detection system that attained an F-score (70.3) greater than results attained for complex one, better-performing method in the competition. From the thorough analysis using an oracle system that has access to aureate sentiment and target of opinion annotations were capable to expect stance with an F-score of 59.6% only. Sriram et al, [18] have explained the classification of text from the Twitter text by using a small set of domain-specific features derived from the author’s profile. The selected approach efficiently categorizes the text into a predefined set of common classes like Events, Opinions, News, Private Messages, and Deals. Developing a novel curse of dimensionality from the existing data sparseness enhances the accuracy of classification by using a small set of features present as a short text. This model presents a better accuracy when compared to the Bags of Words model [19].

2.2 Based on Ensemble Classifier

With the hasty development of the IT field, user-generated content can be expediently posted on the internet.

Wang et al [20] has studied a comparatively of 3 various ensemble methods based on 5 base learners for evaluating the best performance of sentiment classification. From the entire 1200 comparative group investigations, empirical results describe those ensemble methods significantly enhance the effectiveness of distinct base learners for sentiment classification. Amongst the 3 ensemble methods, Random Subspace had shown better effective outcomes. Ensemble methods are used for the testing of the reliability of the text using the uniformly distributed weights and the data is filtered by employing the Bayesian paradigm. Fersini et al, [21] have optimized the ensemble model for obtaining ‘N’ no of possible solutions for acquiring the best accuracy results [22]. This optimized model gives a heuristic capability to evaluate the discriminative marginal improvement that each classifier affords with respect to an assumed ensemble. Xia et al, [23] have performed a comparative ensemble technique for sentiment classification of the text with 2 features with grammatical and word relation features with Naïve Bayes, SVM, and entropy features. This classification is included with ensemble methods consisting of ‘Fixed’, ‘Weighted’, and ‘Meta’ classifier combinations to evaluate the effectiveness of the methods for sentiment analysis. Cornell movie-review corpora [24] is the dataset used for the analysis. In this overall sentiment polarity or subjective rating is derived to know the effectiveness. This ensemble system is an efficient way to associate diverse classification algorithms and feature sets for better classification performance.

2.2 Research Gap

The above discussed prior social media classification and fake news detection methods focus on classifying the fake news using machine learning algorithms. Several fake news detection research works have utilized ensemble classifiers to improve the classification accuracy in the social network further. However, there are several challenges in building accurate fake news predictive models, that includes limited availability of corpora, huge variations in the users' fake news messages when the user unintentionally generates fake news, and subjectivity in the ground truth labels. Hence the idea of combining context-based and Linguistic Inquiry and Word Count Dictionary to provide accuracy in the detection of fake news that can work without social signals. This eases the work of early detection of fake news and in turn, can limit the spread of fake news as a whole. In addition, context-based sentiment analysis is still in its infancy stage in the social network due to the dynamic change of the information context over time.

3.0 MATERIALS AND METHODS

In the previous section, it has been discussed about the Fake news detection and methods used to find out the fake news in any text or from any media that gives information. In the following section, we elaborate on our proposed framework, the algorithm description followed by the Performance Evaluation Metrics. Though there exist multiple challenges in predicting Fake news like a very limited amount of data available, a large amount of time spent by the user over the internet where sometimes un-intentionally or intentionally makes them generate and disseminate the false news.

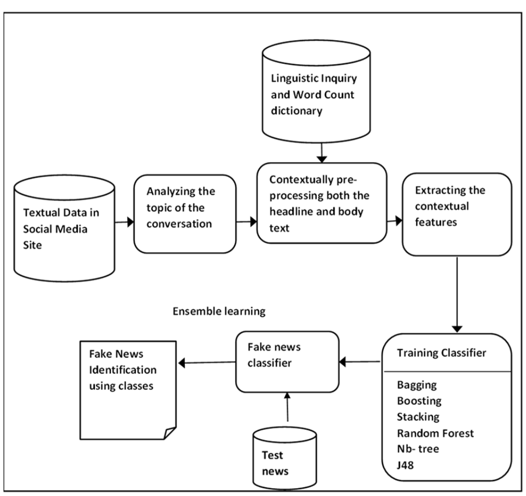


Figure 1. The proposed fake news architecture

This is very dangerous and sometimes times there occurs a chance of spreading a negative impact to society. In fact, defeating such situations there is enormous research work carried on to identify the real news using Ensemble classification algorithms in order to predict with much accuracy. In addition, context-based sentiment analysis is also carried out. However, it is in its infancy stage in the social network due to the dynamic change of the information context over time. Also based on the Linguistic Inquiry and Word Count dictionary, there is no particular research that can assist in the decision-making of the Fake News detection.

This study aims to improve the classification accuracy of fake news detection and to reduce the computation time on a large scale. Hence the work introduces a model with combinations of the Ensemble classifiers using Linguistic Inquiry and Word Count (LIWC) dictionary and CNN model. Figure 1. illustrates the outline process of the proposed fake news detection methodology in the social network. Initially, the proposed approach is targeted to find the discussed topic on a social network through the topic modeling method. After performing the topic modeling, it contextually applies the pre-processing method on the social media messages, including the headline and body of the text with the help of the LIWC dictionary. LIWC dictionary assists to extract the textual data contextually from cognitive, emotional, and spoken categories. The proposed approach extracts the features of the posted or shared content in a social network without modifying their original context.

Hence, it focuses to maintain the textual words in a sentence in its original structure until it is given as the input to the Ensemble learning classifiers. By applying the stacked ensemble classification model, the proposed approach learns the represented text model and classifies the textual data into real news and fake news.

Then, the proposed approach takes the collaborative decision from the classified data that is generated from the multiple base learners using the weight-based Ensemble method. Finally, the proposed approach classifies the real and fake news from the shared news information by exploring both the headline and the body of the text in the content.

The novelty of the proposed research is that the framework implemented has used various Ensemble techniques such as bagging, weight-based ensemble method to increase the performance of the model.

3.1 Ensemble Algorithms

Ensemble methods are meta-algorithms that associate numerous machine learning methods into one analytical modeling order to reduce variance(bagging), bias (boosting), and enhance predictions (stacking). Ensemble learning algorithms are used for training different models to create a final prediction model. The main purpose of ensemble methods is that they are used to reduce the variance, and bias and to improve the prediction accuracy. The work uses supervised machine learning classifiers such as NB tree, J48, Random Forest, and Logistic regression. To train these classifiers three different models have been used for feature extraction. Actually, these features are used to train the classifiers. These models are the stacking, bagging and boosting Model. These models extract the features from the training data set and then the classifier is trained through these features. Ensemble algorithms are popular for better prediction model that combines several prediction models to increase the prediction accuracy. Many ensemble methods are available to build a prediction model. Each model can be different in the way they do behave based on population, technique of modeling, and hypothesis. The important question is how to combine models to classify a better prediction model. It is similar to voting rule, majority of result will decide which algorithm to select. Three popular ensemble methods are stacking, boosting, and bagging. Bagging can be applied for dataset with huge volume of data. Training algorithms are used here to build a prediction model based on various algorithms. Final prediction model is chosen based on majority of voting.

The output of all the machine learning algorithms can be combined to compute the final outcome of the classifier. Assume that if a boosting is

$$F_{T\,(a)=}\ \sum_{i-1}^{T} f1(a) \qquad (1)$$

$$E_{T=}\ \sum_i E\,[F1\,(a_i)\,\alpha h(a_i)\,] \qquad (2)$$

Where Er is the training error for week learners. F1(ai) is the booster classifier, hypothesis for week learner output is denoted as h(a)

3.2 Bagging

Public auditing scheme of cloud assisted WSN-IoT is Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, data is trained by different trees on different subfields (randomly selected by replacement) and calculate the ensemble: Base learners use a boosting model to extract the data sets for training. For acquiring the outcomes of base learners, bagging is used for classification and residuals for regression.

3.3 Boosting

Boosting is another kind of ensemble classifier that is widely used for messy datasets. In particular, these boosting ensemble classifiers are called Meta algorithms. These classifiers reduce the bias and variance. This method boosts the weak learner and makes them strong models. To get precise and accurate results the probabilistic results of ensemble classification are done. Boosting is done sequentially. The given problem uses initially equal-weighted coefficients for all data points which then allows weak learning to correctly classify the data points in an incremental approach. successively the weighted coefficients are decreased for data points that are correctly classified and increased for those which are misclassified [25-28]. More weights are assigned to the weaker ones that have been identified in the earlier study. Finally, it produces the result by combining through weight-based majority vote or by regression. The only drawback that can be faced in boosting is that sometimes the model overfits the data and predicts the incorrect instances

3.4 Stacking

Stacking is the technique used to combine multiple models like classification through a meta-classifier. Every single classification in this model is trained based on the entire training set. Finally, the outputs are classified based on the fitting of the meta classifier. This meta classifier can be either trained well on a predicted label or the probability of the Ensemble classification [29].

Algorithm 1: Stacking
//Input: Training data: Td = i=1 to m{xi, yi}
// Output: Ensemble Classifier E
Set Ensemble learning classifiers
for I = to M do
Learn Ei based on Td
end for
Construct data set for predictions
for j = to n do
Et = {xi, yi }
end for
Learn a meta classifier
Learn M based Eh
Return M

Along with this proposed Ensemble framework, the Learning algorithm NBTree, J48, Logistic Regression, and Random Forest are conjugated to explore the performance evaluation of classifying the fake news using the ensemble classifier.

3.5 Learning Algorithms

3.5.1 NBTree

NBTree is a combination technique of hybridizing the naïve Bayes classifier [30] and decision Tree [31]. This can be implemented over large amount of data and expect the accurate prediction. The trained and learned knowledge is represented in the form of Tree. The construction of the tree is done recursively.

3.5.2 J48

The algorithm is used to create a decision tree which is most appropriately used for Classification and Prediction problems. The tree is generated with multiple nodes and internodes. Decision tree classifiers are particularly used to make some decisions in critical situations. Tress are modeled by splitting up the input data based on previously trained splitting criteria [30]. The representation of the Decision Tree is rather similar to flow charts denoting the instances. Classification is done based on the selected feature values. Nodes in the trees denote the input instance, wherein the outputs are termed as branches and the leaf nodes are the class labels. The J48 classifier is an application of the C4.5 decision tree algorithm[32]. From the given attribute values, the decision tree is developed and classifies the new instance. When a new training set is given, it immediately responds and takes the responsibility to accurately classify the various instances by eliminating the irrelevant and ambiguous data [33-37].

3.6 Logistic Regression

This algorithm is used in classifying the given problem into multiple or binary classes as Yes/No, True/false, Fake/Real and predicts the output in a discrete/ categorical nature. Logistic Regression does not depend on a value to be in range. They use the Sigmoid function curve, where the Sigmoid curve converts any value from negative to infinity or to a discrete value, which actually Logistic Regression works. The sigmoid function curve works as a transition for Logistic Regression because it transforms an output to a probability value. The LR hypothesis function uses a threshold value to indicate the probability of 0 or 1, where it means that the value 0 indicates the predicted output as true and 1 result as Fake.

3.7 Random Forest

The Enhanced Technique of Decision Tree is called Random Forest. It is a collection of Multiple random decision trees and it is less sensitive to the training datasets. They use multiple trees to randomly select the subset feature, hence it is known as random forest. Random forests are modeled to overcome the drawbacks of the Decision tree. The Decision tree has high variance hence tends to overfit the model. Multiple decision trees are built using the bootstrap technique. Where bootstrap technique ensures that they do not use the same data for every tree, in a way it helps the model to be less sensitive to the training data. The random feature selection helps to reduce the correlation between the trees. From the multiple decision trees, the predictions are noted. Finally, the predictions are combined. As it is a classification problem the majority voting is taken. The process of combining results from multiple models is called aggregation. Random forest is the most commonly used technique. As the researchers found the values close to log or the square root of the total number of features works well. For the regression problem, while combining the predictions the average is taken.

3.8 Performance Metrics

To evaluate the performance of the fake news detection algorithm, the experimental framework employs various evaluation metrics. The proposed approach considers the fake news problem as a classification problem that predicts whether a posted social media message is fake or real.

Precision
Recall
F-measure
Classification Accuracy

Precision:

It is the ratio between the number of accurately predicted fake news and the total number of predicted data that are annotated as fake news.

Recall

It is the ratio between the number of accurately predicted fake news and the total number of data that are annotated as the fake news.

F-measure

F-measure or F-score is the harmonic mean of the precision and recall.

Classification Accuracy

It is the ratio between the number of accurately predicted fake news and real news and the total number of data that are in the social media messages

$$\text{Accuracy} = \frac{\text{TRUE POSITVE} + \text{TRUE NEGATIVE}}{100}$$

$$\text{Precision} = \frac{\text{TRUE POSITVE}}{TRUE\ POSITIVE + FALSE\ POSITIVE}$$

$$\text{ROC} = \frac{\text{TRUE POSITVE RATE}}{FALSE\ POSITIVE\ RATE}$$

$$\text{Sensitivity} = \frac{\text{TRUE POSITVE}}{TRUE\ POSITIVE + FALSE\ NEGATIVE}$$

$$\text{Specificity} = \frac{\text{TRUE NEGAITVE}}{FALSE\ POSITIVE + TRUE\ NEGATIVE}$$

4.0 EXPERIMENTAL ANALYSIS

4.1Formulation of Dataset

We used ISOT dataset and publicly available datasets in Kaggle. ISOT dataset consists of both fake news article and true news articles. Both types of articles are collected from world wide web. Original article are collected from reuter news website and false news dataset in ISOT are collected from different open sources, most of them are flagged by Politfact. Out of 45,000 news articles, 21499 are true articles and 23501 are false articles. Most of the false articles were targeting political news. A combined dataset is used in the work to identify the performance of the proposed algorithm.

4.2 Formulation of Feature Sets

Many researches have been emerged for fake news detection. The main aim of this study is to distinguish fake news and true news from the available features. Initially, the data set is collected from the source in the string format and the data is read by the CSV file. Two CSV files such as True.csv and Fake.csv were used in the dataset. The above-discussed algorithms are applied to detect fake news with the help of JDK, and NetBeans with a Xamp server from a data set. The work uses Kaggle dataset where real news articles are collected from reuters and many unreliable sources flagged by Politifact as part of fake news articles. The dataset consists of four features such as id: identity of a news article, title: title of a news article, author: author name of the news article, text: the text of the article; not complete, the target is “label” that consist of binary digits 0s & 1s. Where 0 indicates reliable source of news (not fake news). 1 indicates fake news. The dataset is separated in to train and test to check the accuracy through training and testing on it.

4.3 Results

The experiment was conducted with the help of Python environment. Pyton-weka-wrapper3, Numpy, Pandas, NLTK, Keras are the python library files to conduct this experiment. The performance evaluation for the proposed work is measured in terms of accuracy, precision, Recall, and F-measure. From the input of the data, the topic is analyzed based on a probability model. In this analysis, the data is filtered using the Latent semantics of the text. This semantics is analyzed using the vocabulary and words in the dictionary, from that analysis the word index is identified. The latent semantics helps in identifying either the similar or the related terms to the target keywords. In the next phase, two options were given in order to collect the database. Primarily the benchmark dataset can be taken or else generating the source file directly from the link. The input source is selected from the link and the text is read from the source. From the text source, the Pre-processing of the text is done through tokenization, NLP, Linguistic Dictionary and form the words based on vocabulary. The text is analyzed in the next phase. The Processing is carried out to check the trueness of data.

The data is analyzed from the source and divided into sequences of the strings, framing them into words, keywords, phrases, symbols and, then the word count is calculated. In this pre-processing stage, the words analyzed from the NLP Linguistic Dictionary are tokenized, thus converted to text and used for further extraction. From the pre-processing phase, the text is analyzed to extract the features using the NLP technique with LIWC dictionary based on the POS model with whitespace Tokenizer. During pre-processing stage, all the features are trained using machine learning classifiers. Then, we used 20-fold cross validation to divide the data into 20 portions. K-fold cross-validation will effectively use the dataset for the both testing and training. First, we calculate the error rates Average Absolute Error (AAE) and Average Relative Error (ARE) as shown in table 1. Error rates calculation are done based on the following formula as in equation 3 & 4:

$$AAE = (^1/_n) \sum_{i=1}^{n} |\ (Pi - P)| \tag{3}$$

Where Pi represent the value of prediction and P represent the actual value and n is the number of parameters used in the calculation.

ARE is used to measure the whole size of the object with absolute error.

$$ARE = (^1/_n) \sum_{i=1}^{n} |\ (Pi - P)|\ /\ (Pi + 1) \tag{4}$$

Table 1 Error Rate Analysis

Techniques	AAE	ARE
Boosting	0.045607	0.043342
Bagging	0.055312	0.052132
Random Forest	0.038301	0.033287
Logistic Regression	0.047607	0.043402
J48	0.055622	0.053347
Boosting	0.045607	0.043342
NB tree	0.052271	0.052376

The evaluation of the proposed work indicates the better performance of the model with other models. Most of the model utilizes hybrid approach by combining two or more models for improved detection accuracy [38]. The reason for performance improvement is due to the efficient feature selection with the help of machine learning models. The ensemble technique along with the Linguistic Inquiry Word Count dictionary is the novelty of this work. To generate the best feature from the corpus, LIWC 2015 is incorporated. The LIWC dictionary extracts various kinds of linguistic features. Here, the string is mapped with 5 different categories as Social, Affective, Cognitive, Perceptual, and Biological. Few to mention are positive emotions, words indicating Negative Emotions, stop words, Function words, Punctuations used, the informal language used, certain grammars used in the sentences such as articles, preposition, adjectives, and adverbs. From the extracted text the dependency probability is performed as textual analysis using the NLP technique. From this dependency probability, the text is classified into a set of data. The predicted features are then used to train the machine learning model, using the Ensemble classifier which is the second novelty of this work to find the performance and accuracy.

The evaluation result of table 2 shows that the proposed model performs best with precision, recall and F-measure. Certain measures have been taken to ensure that the model does not overfit or underfit the data. In the classification phase, the data is sent to the classifier to give the Choice of the prediction using Ensemble methods.

Table 2. Evaluation of proposed model with ensemble classifiers

Classifier	Precision	Recall	F-measure	Accuracy
Nb-tree	86	91	0.53	77
J48	84	92	0.53	86
Logistic regression	84	88	0.54	84
Random forest	85	95	0.96	97
Bagging	74	85	0.74	92
Boosting	72	95	0.86	94
Nb-tree	86	91	0.53	77

The work uses three learning models: bagging, boosting, and Stacking. Bagging and Boosting are based on voting classifier which consist of RF, LR, J-48, and NB-tree algorithms. Voting model can be used in classification to allow two or more models for the whole dataset. Voting classifiers in the work are RF, LR, J48 and NB tree. Stacking will predict the features through meta classifier. First the proposed model is trained against voting classifier. Then we test the model using error rate analysis based on the majority of votes by all these algorithms. From the choice of algorithm, the ensemble classifier does the testing and predicts the result based on the majority of votes given by the models. The work uses different metrics to perform the evaluation of the proposed work. Most of the metric are based on confusion matrix as shown in table 3.

Thus, the proposed model yields the better performance in time, Accuracy, and Prediction from the chosen fake news social site datasets. Using the confusion matrix with the given total number of instances, the classified data is analyzed and generates the results of the performance with the performance time and the state of accuracy. Then the data is predicted and presented as mostly false, mostly true and the mixture of true and false.

Table3. Overall confusion matrix

	Positive	Negative
True	11915	520
False	1397	2449

4.4 Accuracy Prediction

From the entire classification and detection, the accuracies of the data sets are measured as per the number of records been scanned (Table 4). The runtime of the same records is measured in n number of times of scanning. So, the runtime is termed as choices 1, 2, and 3. The accuracy of scanning the data is shown as increase from the first time to the second time (i.e., ES_Choice I to ES_Choice II), and from second time to third-time accuracy is decreased (i.e., ES_Choice II to ES_Choice III) as number of records increases, the scanning of the records shows the best percentage of accurate prediction in ES_Choice II

Table 4 Ensemble classification

Time taken to build the model	11 Seconds
Correctly Classified Instances	14364
Incorrectly Classified Instances	1917
Kappa statistic	0.6458
Mean absolute error	0.1177
Root mean squared error	0.3431
Relative absolute error	32.6285%
Root relative squared error	80.7838%
Total Number of Instances	16281

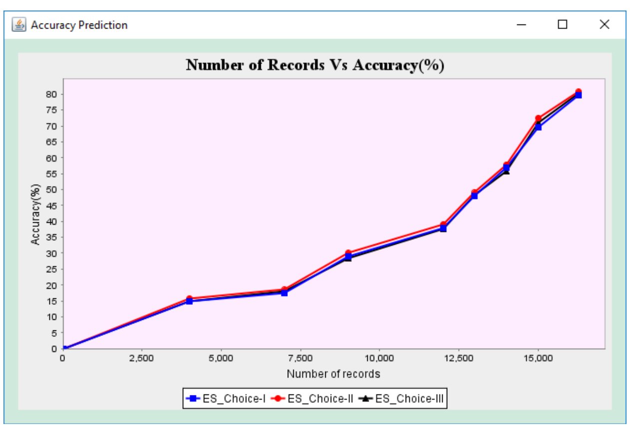


Figure 2 Accuracy Prediction

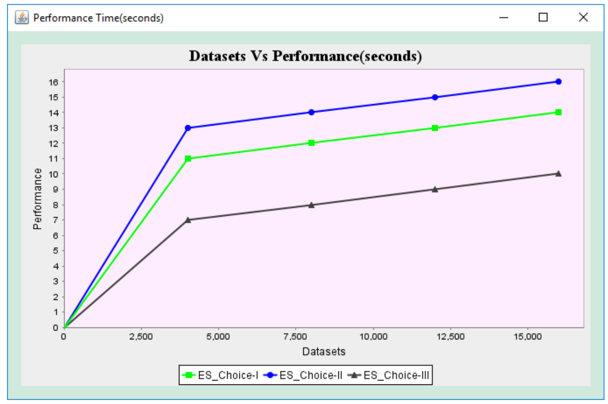


Figure 3 Performance Evaluation of Time(Seconds)

4.5 Performance Evaluation

As similar to the accuracy prediction, the performance of the algorithm is also analyzed. From Figure 2. it is observed that the performance of the choice II gives the best performance in prediction accuracy.

Figure 3 shows the performance evaluation of time consumption of the algorithm. As the number of the records, increases the accuracy prediction and performance of the algorithm increases, but it is limited to the 2 runtimes of the algorithm to get the best performance and high percentage of accurate prediction.

4.6 Comparison

The obtained result is compared with the single classifier to exhibit the performance and the error rate. Here, the Accuracy is validated between the Ensemble classification method and a single classifier Naive Bayes classification method. As per the expectation, our proposed Ensemble method outperforms more accurately than the Existing Naïve Bayes shown in Figure 4. It has been proved that the combination of works done by the expertise sounds better compared with a single men army. It performs less when used as a single classifier, rather combining with the other classifier random forest, J48, Logistic Regression as an ensemble shows better performance. Figure 5 shows the error rate analysis. Figure 6 shows the time spend by the algorithm in prediction analysis.

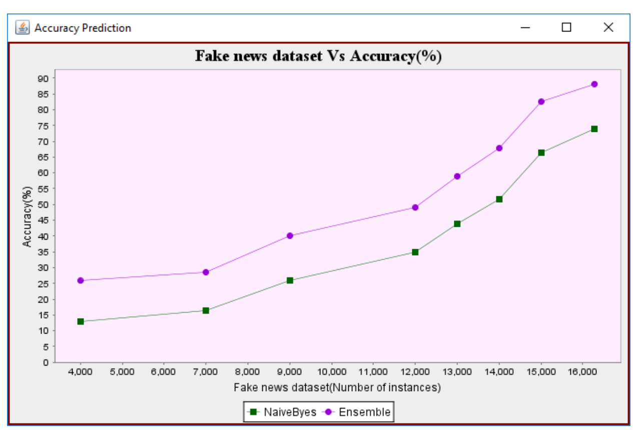


Figure 4. Accuracy Prediction

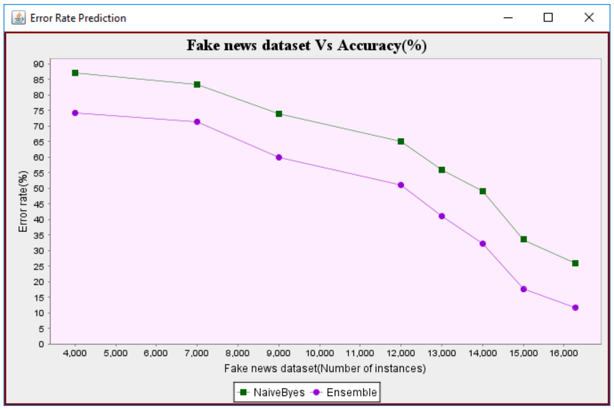


Figure 5. Error Rate between Ensemble Method and Naive Bayes

Finally, the Performance time was validated between Ensemble classification Method and Naive Bayes classification but the proposed Ensemble method is efficient compare to Existing Naïve Bayes.

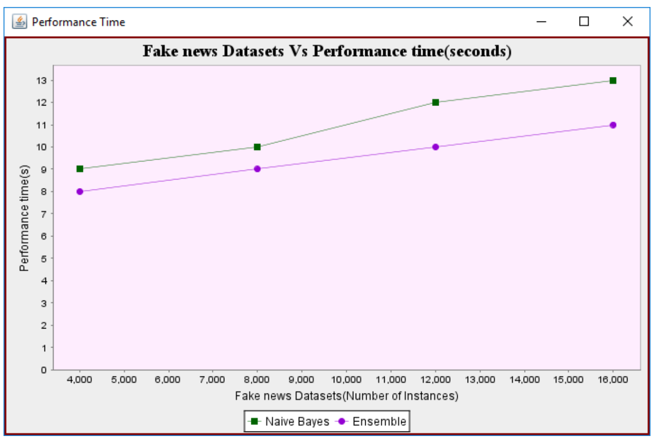


Figure 6. Performance Time Prediction

5.0 CONCLUSION

The main aim of the study is to eradicate the drawbacks of social media, which is considered as a great tool for spreading fake news. With the growth of the vast amount of user-generated data in the social network, the social network necessitates an accurate method to reduce the information overload as well as to avoid the fake news spreading over the world. Hence, this work presents a notion for a Fake news detection model in an online social network with the help of an Ensemble Classifier. To obtain the objective of fake news detection model, the work targets to identify the topic model and exploit the LIWC dictionary to contextually classify the Fake news from the abundant social media messages. The proposed approach applies a Stacked Ensemble Classifier that consists of a set of base learners to recognize the exact stance based on the decision-making. Different textual properties were explored to differentiate fake news and real news. The work also uses multiple learning algorithms for training and the performance were evaluated with the help of the open-source dataset. The accuracy prediction and performance evaluation of time consumption of detecting Fake news is 97% using random forest algorithm and 11 ms respectively. Thus, it improves the classification accuracy over large-scale social media messages through an efficient Sentiment Analysis model. As part of future work, we plan to combine several machine learning algorithms to consider the image features.

Acknowledgement

We would like to thank Periyar University as well as Periyar University College of Arts and Science for providing valuable insights and support for the conduction of this study.

References

[1] MontherAldwairi, AliAlwahedi. 2018. Detecting Fake News in Social Media Networks. *Procedia Computer Science*, 141: 215-222.

[2] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. 2017) Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1): 22-36.

[3] Allcott, H., & Gentzkow, M. 2017. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2): 211-36.

[4] Chen, Y., Conroy, N. J., & Rubin, V. L. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* 15-19. ACM.

[5] Hušek, M. 2018. *Text mining in social network analysi*s. Charles University.

[6] Chen, Y., Conroy, N. J., & Rubin, V. L. 2015. News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, 52(1): 1-4.

[7] Freeman, D. M. 2017. Can you spot the fakes?: On the limitations of user feedback in online social networks. In Proceedings of the *26th International Conference on World Wide Web.* 1093-1102. International World Wide Web Conferences Steering Committee.

[8] Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. 2018, Fake News Detection in Social Networks via Crowd Signals. In Companion of the The W*eb Conference 2018 on The Web Conference 2018* 517-524. International World Wide Web Conferences Steering Committee.

[9] Cao, Q., Sirivianos, M., Yang, X., & Pregueiro, T. 2012. Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. 15-15. USENIX Association.

[10] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. 2018. Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR), 51(2): 32.

[11] Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., ... & Havlin, S. 2018. Fake news propagate differently from real news even at early stages of spreading. arXiv preprint arXiv:1803.03443.

[12] Chen, Y., Conroy, N. J., & Rubin, V. L. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* 15-19. ACM.

[13] Rubin, V. L., Chen, Y., & Conroy, N. J. 2015. Deception detection for news: three types of fakes. In Proceedings of the *78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*.83. American Society for Information Science.

[14] Y. Shukla, N. Yadav, and A. Hari, 2019. "An unique approach for detection of fake news using machine learning," *Proceedings of the International Journal for Research in Applied Science and Engineering Technology,* 7(6): 491–496.

[15] Anu Shrestha, 2020 Francesca Spezzano, and Abishai Joy, "Detecting Fake News Spreaders in Social Networks via Linguistic and Personality Features Notebook for PAN at CLEF", *CEUR Workshop Proceedings*

[16] Lex, E., Juffinger, A., & Granitzer, M. 2010. Objectivity classification in online media. In Proceedings of the *21st ACM conference on Hypertext and hypermedia.* 293-294. ACM.

[17] Zhang, J., Cui, L., Fu, Y., & Gouza, F. B. 2018. Fake News Detection with Deep Diffusive Network Model. arXiv preprint arXiv:1805.08751.

[18] Mohammad, S. M., Sobhani, P., & Kiritchenko, S. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT),* 17(3): 26.

[19] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 841-842. ACM.

[20] Zhang, Y., Jin, R., & Zhou, Z. H. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4): 43-52.

[21] Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. 2014. Sentiment classification: The contribution of ensemble learning. *Decision support systems,* 57: 77-93.

[22] Fersini, E., Messina, E., & Pozzi, F. A. 2014. Sentiment analysis: Bayesian ensemble learning. *Decision support systems*, 68: 26-38.

[23] Partalas, I., Tsoumakas, G., & Vlahavas, I. 2010. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning,* 81(3): 257-282.

[24] Xia, R., Zong, C., & Li, S. 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6): 1138-1152.

[25] Pang, B., & Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the *42nd annual meeting on Association for Computational Linguistics.* 271. Association for Computational Linguistics.

[26] Van Puyvelde, D., Coulthart, S., & Hossain, M. S. 2017. Beyond the buzzword: big data and national security decision-making. *International Affairs,* 93(6): 1397-1416.

[27] Pat Langley, Wayne Iba, Kevin Thompson.1992. An analysis of bayesian classifiers. *National Conference on Artificial Intelligence*, 223–228.

[28] Ross R. Quinlan.1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.

[29] Mahabub, A. A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. SN Appl. Sci. 2, 525 (2020). https://doi.org/10.1007/s42452-020-2326-y

[30] Yuan D, Lu X, Li D, Liang Y, Zhang X. 2019. Particle filter re-detection for visual tracking via correlation filters. *Multimedia Tools and Applications* 78(11):14277–14301

[31] X. Zhou, R. Zafarani, K. Shu, and H. Liu, 2019. “Fake news: fundamental theories, detection strategies and challenges,” in Proceedings of the *12th ACM International Conference on Web Search and Data Mining, Melbourne, Australia,* January

[32] D. de Beer and M. Matthee, “Approaches to identify fake news: a systematic literature review,” in Proceedings of the International Conference on Integrated Science, Kep, Cambodia, May 2020.

[33] Hasan, H., & Hashim, L. 2009. What's new in online news. PACIS 2009 Proceedings, 1-13, Hyderabad, India

[34] A. Bondielli and F. Marcelloni, 2019. “A survey on fake news and rumour detection techniques,” *Information Sciences*, 497: 38–55,

[35] Gutierrez-Espinoza, F. Abri, A. S. Namin, K. S. Jones, and D. R. Sears, 2020. “Fake reviews detection through ensemble learning,” https://arxiv.org/abs/2006.07912, Cornell University

[36] Y. J. Lu and C. T. Li, “GCAN: graph-aware co-attention networks for explainable fake news detection on social media,” https://arxiv.org/abs/2004.11648, 2020.

[37] Auhtor? 2020, https://www.kaggle.com/mrisdal/fake-news Fake or Real News Retrieved on 11 June 2023

[38] Adil H Khan,Dayang Nur Fatimah Awang Iskandar, Jawad F. Al-Asad,Hiren Mewada,Muhammad Abid Shera. 2022. *Ensemble learning of deep learning and traditional machine learning approaches for skin lesion segmentation and classification*. *Concurrency and Computation Practice and Experience* 34(5) https://doi.org/10.1002/cpe.6907