Machine Learning Fake News Classification with Optimal Feature Selection

Muhammad Fayaz
University of Peshawar
Atif Khan (atifkhan@icp.edu.pk)
Islamia College Peshawar https://orcid.org/0000-0003-3628-0262
Muhammad Bilal
Islamia College Peshawar
Sanaullah Khan
Kohat University of Science and Technology

Research Article

Keywords: Machine Learning, Random Forest, Fake news, Feature Selection.

Posted Date: September 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-835344/v1

Version of Record: A version of this preprint was published at Soft Computing on January 29th, 2022. See the published version at https://doi.org/10.1007/s00500-022-06773-x.

Abstract

Nowadays, information is published in newspapers and social media while transmitted on radio and television about current events and specific fields of interest nationwide and abroad. It becomes difficult to explicit what is real and what is fake due to the explosive growth of online content. As a result, fake news has become epidemic and immensely challenging to analyze fake news to be verified by the producers in the form of data process outlets not to mislead the people. Indeed, it is a big challenge to the government and public to debate the situation depending on case to case. For the purpose several websites were developed for this purpose to classify the news as either real or fake depending on the website logic and algorithm. A mechanism has to be taken on fact-checking rumors and statements, particularly those that get thousands of views and likes before being debunked and refuted by expert sources. Various machine learning techniques have been used to detect and correctly classified of fake news. However, these approaches are restricted in terms of accuracy. This study has applied a Random Forest (RF) classifier to predict fake or real news. For this prpose, twenty-three (23) textual features are extracted from ISOT Fake News Dataset. Four best feature selection techniques like Chi2, Univariate, information gain and Feature importance are used for selecting fourteen best features out of twenty-three. The proposed model and other benchmark techniques are evaluated on the dataset by using best features. Experimental findings show that, the proposed model outperformed state-of-the-art machine learning techniques such as GBM, XGBoost and Ada Boost Regression Model in terms of classification accuracy.

1. Introduction

The internet offers many possibilities along with many challenges when it comes to reporting the news. The number of communication channels is growing over time. In addition to conventional channels such as newspapers and TV channels, news communication channels such as blogs and social networks have arisen since the internet became a spreading source. It has become simpler for customers to receive the latest news on their fingertips. In the current state, these social media sites are highly effective and valuable if this change has a positive aspect on one-hand while negative aspect on the other hand such as fake and inaccurate news because editorial boards do not necessarily determine the trustworthiness of the information posted. In the current state, these social media platforms are useful to share ideas and discuss issues such as governance, education and health. Organizations widely use most sites for their monetary benefits for their objectives.

A study placed by Twitter reveals that fake news is 100 times speedily spread rather than real news. The phenomenon of fake news can produce effects that are not significant, and this can lead to results that affect millions of people in certain countries (Vogel and Meghana 2020). Such a propanganda and rumors fluctutate stock prices, stock purchases, investment plans and even reaction to natural disasters. The contents of fake news are frame in such a way that it may create mass opinions and fully win over the reader to make them completely confused and their attention divert from real news (Hakak et al. 2021). Detection of fake news is a challenging task, as that requires rationalism. Many fact checking

websites are deployed to reveal the fake news to counter the growing misinformation. Such websites play a critical role in clarifying false news, but they need time-consuming and expertise. It is quite challenging to detect and analyze the data authenticity(Napoli 2018).

This study uses benchmark and other machine learning approaches for fake news classification. This research uses Random Forest (RF), as proposed machine learning algorithm, to improve the accuracy of fake news classification. The proposed machine learning model operates as follows: first we extract twenty-three (23) textual features from the ISOT Fake News Dataset publicly available on the ISOT website and describe the news dataset as a features vector. Too many features influence model efficiency and performance, and not all features have the same predictive model contribution. So it is important to strip out non-valuable and less important features to reduce model complexity and increase model accuracy. Therefore, four different feature selection techniques are used, such as Chi-Square, Feature Importance, Information Gain and Univariate, to pick the fourteen (14) best features out of 23 features. Based on the above, we will then abstract the real information from the fake news dataset. For this analysis, the efficiency of the proposed model is contrasted with the benchmark techniques.

This study's contributions are as follows:

- To propose a Model (Random Forest) for the classification of fake and real news.
- To test the proposed classification model for all textual features (twenty-three) derived from ISOT fake news dataset.
- To determine the efficacy of the proposed model with respect to the best features collected through four different feature selection techniques (Chi-Square, Univariate, Feature importance, Information Gain).

Remaining portions of the paper are designed as follows: Sect. 2, comprehensive literature is presented. Next Sect. 3, the proposed machine learning model is illuminated. Section 4, various experimental findings and a complete discussion is shown. Final Sect. 5, the conclusion and future recommendation.

2. Related Work

The news are very important because it keeps the public informed about activities and events around their premises and beyond their premises. Reports showed that most adults use digital forms such as social media and web/search engines to access their news instead of using traditional media. Fake news detection has got considerably attention (De Choudhury et al. 2014). In this section, numerous methods have been suggested to detect fake news in various types of features and datasets. Authors in (Okoro et al. 2018) used a Machine-Human (MH) model for detection of fake news on social media. The study (Khan et al. 2021) focused on detectition of fabricated opinions and compared the Glove Embedding and Character Embedding features of fake and real news dataset using three datasets. Amongt them, two are standard datasets and one is combination news of distributed topic on social media through Naïve Bayes, CNN, LSTM, Bi-LSTM, C-LSTM, Heterogeneous Graph Neural Network (HAN), Cov-HAS, Char-level

C-LSTM model. It is observed that n-gram features show great promising results in fake news on Naive Bayes model which is almost equivalent to the performances of CNN based model. The authors in (Ozbay and Alatas 2020) used mixure of text classification techniques and supervised artificial intelligence classifiers. The proposed model was tested on three different real word datasets. The model was evaluated using accuracy, precision, recall and F-measures values. The performance of best mean values was obtained from the Decision Tree Algorithm. Zero, CV parameter selection (CVPS) with 1000 value, seems the best recall metric algorithm. Authors in (Gravanis et al. 2019) used an enhanced set of linguistic features for the detection of fake news by evaluated several classification models using five different datasets containing fake and reals news. Adaboost obtained 95% accuracy over all datasets and next is ranking Support Vector Machine (SVM) and Bagging algorithms.

The study (Ahmad et al. 2020) presented work of machine learning model and ensemble techniques for detection of fake and real social News. Data collected from web contains fake and real news covering different domain. They extracted different textual features from the dataset and used it as input to different machine learning models like Logistic Regression, SVM, MLP, KNN, Random forest(RF) and ensemble models like voting classifier(RF, LR, KNN), voting classifier(LR, LSVM, CART), Bagging classifier (decision tree) and boosting classifier (AdaBoost and XG-Boost), Ensemble model XGBoost performance better than other classifiers and ensemble model in terms of accuracy. The author in (Ahmed et al. 2019) have used n-grams and Part of Speech (POS) tagging, they suggested Deep Syntax Analysis using Probabilistic Context-Free Grammars (PCFG). The author in (Ruchansky et al. 2017) proposes the CSI hybrid model used for fake news detection. The CSI model is comprised of three modules. The first model captures the pattern of the user's temporal engagement with an article. The second modules capture the characteristic source present in the behavior of users and the third modules are used as integrated of both modules first and second experiment on two datasets, check robustness of CSI model when labeled data is limited. It also inspects suspicious users' behaviors. The CSI model doest not make assumptions regarding distribution of user behavior specially textual context of the data or the structure of data underlying. In the study of (Mansouri et al. 2020), a combined method based on semi-supervised LDA (Linear Discriminant Analysis) and convolutional neural network are used to detect fake news using an unlabeled dataset for the convolutional neural network the unlabeled dataset is labeled. The result of the proposed method of precision is 95.6% and 96.7% recall, which outperforms existing methods for detecting fake news.

Another study (Najar et al. 2019) Fake news detection using Bayesian interference used Bag of Words using Multinomial Model (MM), Dirichlet Compound Multinomial (DCM) and Deterministic Annealing Expectation-Maximization(EDCM-DAEM) and EDCM-Bayesian, EDCM-Bayesian better accuracy than other classifiers, classification accuracy 87.85 on BS-Detector dataset. In study (Jain et al. 2019; Reis et al. 2019) used different textual features like language features (syntax) such as n-gram and part of speech tagging, lexical features (character and word-level signals), psycholinguistic features, semantic features and subjectivity and sentiment scores of a text using classification of K-Nearest neighbors (KNN), Naïve Bayes(NB), Random forests(RF), Support Vector Machine (SVM) with RBF kernel (SVM), and XGBoost (XGB). Random forest and XGB performed best using handcraft features, web-based networking

media. In study [14] using Naïve Bayes classifier, SVM with comparison Naïve Bayes and CNN. Results show that Naïve Bayes, SVM, NLP are performed better than other machine classifier.

The accuracy of proposed model 93.50% at the other machine learning model.

Another study (Faustini and Covões 2019) conduct on fake social media news used three datasets of social media (Twitter, WhatsApp and Fake BR Corpus), by extracting of fourteen textual features such as proportion of uppercase characters, exclamation marks, question marks, number of unique words, number of sentence, number of characters, words per sentence, proportion of adjective, adverb, nouns, sentiment of message, proportion of swear words and proportion of spell errors as features for classifiers. In study (Hlaing and Kham 2020) presents multidimensional fake news (news content, social engagement and news stance) used synonym-based features using three different classifier Decision Tree classifiers, AdaBoost classifier and Random forest classifier for detection of fake news. Experimental result show that Random Forest perform better than other two classifiers on social media dataset.

The study (Mahir et al. 2019) reported that SVM performed better than other classifiers including Naïve Bayesian, RNN/LSTM, Logistic Regression in recognizing fake news extracted from twitter. The study (Al-Ash et al. 2019) used Indonesian news dataset consist of fake and real news documents to show the classification performance of Random Forest, SVM and Naïve Bayesian Classifiers over this dataset with associate classification approach. In this study (Katsaros et al. 2019) eight models were evaluated for classification purpose. These models include Linear Regres. sion, SVM, MLP, Gaussian and Multinomial naïve Bayes, Random Forests, Decision Trees and CNN on three publicly available datasets. The result showed that the CNN is the best performing algorithm.

The study of (Choudhary et al. 2021) proposed a deep learning architecture called BerConvoNet for classification of Fake news and Real news with marginal error. The proposed architecture was composed of two main blocks, a New Embedding Block (NEB)and a Multi-scale Features Block (MSFB). The NEB used BERT for extracting word embeddings from news articles and then fed it to MSFB as input.

In the study (Vogel and Meghana 2020) reported that SVM achieved the highest accuracy of 92% in classification of fake news. He used hand crafted features extracted from news dataset like total word(tokens), Unique words, Unique words (types), type/token ratio, Number of sentences, average sentence length, Number of Characters, Average word length, nouns, prepositions, adjectives.The classification models include XG Boost, Random Forest, Naïve Bayesian, KNN, Decision Tree and SVM were used for classification of fake news.

3. Proposed Methodology

The methodology section presents the architecture of proposed model for classification of fake news as shown in Figure-1. The methodology section is consists of three Main phases: 1st phase: pre-processing (Sentence segmentation, tokenization, stopword removal and word stemming), 2nd phases include

features extraction and best features selection, best feature selection using famous feature selection techniques (Chi-Square, Feature Importance, Information Gain and Univariate) and final phase classification of fake news using the proposed model and other machine learning models.

3.1. Dataset

The proposed machine learning model is tested on the ISOT News Dataset, a publicly available dataset containing false and real news. This dataset is commonly used in the false news identification problem. A total of 44,919 fake and real news was used in this research, including 23502 fake news and 2147 real news assessments using multiple machine learning models. Regarding performance metrics, i.e. classification precision, for the assignment of fake news classification, we also tested the suggested Machine Learning classifier with the benchmark models.

3.2. Pre-processing:

In order to avoid overfitting, the data is preprocessed before fetching it to the Natural Languange Processing (NLP) system. The preprocessing involves various steps like sentence segmentation, tokenization, stopwords removal and word stemming, as discussed below in detail:

3.2.1. Sentence Segmentation:

Sentence segmentation is establish text borders and break the text into sentences. Exclamation (!), interrogation (?), and utter stop (.) signs are widely used as markers to segment the paragraph into sentences.

3.2.2. Tokenization

At this stage, the phrases and sentences are divided into separate words by dividing them into white spaces such as tabs, blanks, and signs of punctuation, i.e. dot (.), comma (,), semicolon (;), colon (:), etc. These are the key indications for dividing the sentences into tokens.

3.2.3. Stopwords removal.

Words which have occurred repeatedly are called stopwords in a sentence. These consist of prepositions (in, on, at, etc.), conjunctions (and thus, too, etc.), articles (a, an, a), etc. These words have little meaning in text documents and are more weighted, and removing them will help improve the system's performance.

3.2.3. Word Stemming:

Stemming is used to bring the word to its basic form. Word stemming plays a significant role in preprocessing. In order to normalize the word token to a standard form, this step changes the derived words to its base or stem word. The famous stemming algorithm, Porter’s stemming (Porter, 1980), is

adopted to remove suffixes like –ing, -es, -ers from the text words. For example, the words ‘looking’ and ‘looks’ will be modified to its base type ‘look’ after stemming.

3.3 Features Extraction:

In text classification problems, features play a major role. This step aims to mine ISOT fake news Dataset features for the problem of text classification. In this study, we extracted twenty-three (23) features form ISOT fake news dataset. Almost all of these features are textual and can be accurately extracted through text as seen in Table 1.

Table 1
List of all features extracted from ISOT News dataset

S.No	Features	Description
1	Counts-words	Count total numbers of words
2	Upper-case-character	Count total number of upper case characters
3	Lowe-case-character	Count total number of Lower case character
4	Character-count	Count total number of characters
5	Count-sentence	Count total number of sentences in fake News
6	Automated-Readability-Index	Automated Readability Index is readability test
7	Coleman-Liau-index	Coleman–Liau index is text readability test
8	Flesch-reading-ease	Flesch reading ease is text readability test
9	Flesch-Kincaid-grade	Flesch Kincaid grade level is text readability test
10	Dale-Chall-formula	Dale–Chall formula is text readability test
11	Topic	Topic related words
12	Count-spaces	Count number of spaces
13	Part-of-Speech	Count numbers of Part of speech
14	Principal-Component-Analysis	Principal Component Analysis
15	Gunning-Fox-index	Gunning fog index is readability test
16	SMOG-formula	Mc-Laughlin's SMOG formula is readability test
17	Sentiment-analysis	Sentiment analysis show text positivity and negativity
18	Bag-of-word	Bag of words is a text representation representing the presence of words
19	Tf-idf (unigram)	TF-IDF with unigram
20	Tf-idf (bi- gram)	TF-IDF with Bi-gram
21	Tf-idf (tri- gram)	TF-IDF with Tri-Gram
22	Stopwords	Count total numbers of stop words
23	Negative-words	Count total numbers of negative word

3.3.1 Features Selection

It is normally not good to use all twenty-three (23) features to classify the ISOT News dataset as fake and real News. All features do not have the same significance and weight when developing a consistent and effective statistical model. Some features are useful and add more to the model prediction and play a vital role in classification accuracy, while others are less valuable and have a less significant impact on the performance of the model. In addition, the appropriate and useful features eliminate over-fitting, increase precision and reduce the predictive model training time. We used the four best features selection techniques to resolve this issue, which are Chi2, Univariate, information gain and Feature importance to decrease the space size of the features to achieve optimum features and significant features. In Table 2, Column-2 shows fourteen (14) most important features for the News dataset were chosen by the Chi-square technique and same numbers of feature selected using Univariate feature selection techniques as seen in Column-3 of Table 2. Similarly, Column-4 picked the fourteen best features from the same dataset using feature importance and Column-5 shows the fourteen best features selected using information gain. Next section evaluated performance of proposed model and other machine learning model on all textual features and best features selected by best features selection techniques as mentioned in Table 1 and Column 2–5 of Table 2.

Table 2
List of fourteen best features chosen by different Features Selection Techniques for ISOT News dataset

S.No	Features Selected by			
	Chi2	Univariate	Feature Importance	Information gain
1	Counts-words	Counts-words	Counts-words	Count-unique-words
2	Upper-case-character	Upper-case-character	Upper-case-character	Upper case character
3	Lowe-case-character	Lowe-case-character	Count-sentence	Count sentence
4	Character-count	Character-count	Flesch-Kincaid-grade-level	Flesch Kincaid grade level
5	Count-sentence	Count-sentence	Flesch-reading-ease	Flesch reading ease
6	Automated-readability-index	Automated-readability-index	Dale-Chall-formula	Dale–Chall formula
7	Coleman-liau-index	Coleman-liau-index	Gunning-fog-index	Gunning fog index is readability
8	Flesch-kincaid-grade-level	Flesch-kincaid-grade-level	Somgs-Formula	SMOG formula
9	Flesch-reading-ease	Dale–Chall-formula	Topic	Count lowercase characters
10	Principal-component-analysis	Topic	Tf-idf (unigram)	Tf_idf (unigram)
11	Tf-idf (unigram)	Count-spaces	Stopwords	Tf-idf(bigram)
12	Tf-idf (bi-gram)	Part-of-Speech	Negative-words	Tf-idf(trigram)
13	Tf-idf (tri- gram)	Tf-idf (unigram)	Part-of-Speech	Stopwords
14	Dale–chall-formula	Sentiment-analysis	Sentiment-analysis	Negative words

3.4 Classification Model for Fake News dataset

This section aims to classify ISOT news as fake and real news using Random Forest and other Machine Learning Model. We train and evaluate the classifiers initially using 10 fold-cross validations on the ISOT False News dataset to validate the impact of the individual models and the proposed model on all textual features and best features selected by features selection techniques. Random forest is a a traditional machine learning algorithm which is used for classification as well as regression problems by using ensembling of many Decision Trees to solve a complex problems. It uses bagging and bootstrap

methods for the prediction of model accuracy. The prediction of each decision trees combines for final prediction using a majority of vote as shown in Fig. 2.

For Decision Trees, Gini Impurity and Entropy are calculated by using following Eq. 1–2, respectively.

$$\text{GiniImpurity} = \sum_{i=1}^{c} fi(1 - fi) \quad (1)$$

$$\text{Entropy} = \sum_{i=1}^{c} - flog(fi) \quad (2)$$

4. Experimental Setting

The proposed machine learning model is tested on the ISOT News Dataset containing a total of 44,919 fake and real news including 23502 fake news and 2147 real news. The dataset is pre-processed by breaking the news text into sentences. The sentences are tokenized into terms and stopwords are removed. Initially, twenty three (23) textual features are selected from the ISOT News dataset for fake news classification. The proposed model and other machine learning models are evaluated on all twenty three features. As all features do not have the same importance in creating a consistent and accurate predictive model. Some features are meaningful and contribute more to model accuracy, while others are less important and adversely affect the model's performance. In addition, the appropriate and useful features eliminate over-fitting, increase precision and reduce the predictive model training time. We used four best features selection technique to resolve this issue, which are Chi2, Univariate, information gain and Feature importance are used for selecting fourteen best features out of twenty-three and then evaluated proposed model and benchmark techniques in terms of accuracy which is used as

performance metrics.

5. Results And Discussions

In the first step, the performance of individual models and other machine learning models are evaluated by using all twenty three (23) textual features extracted from ISOT News dataset. The results of this experiment is recorded in Table 3. It is revealed from the results that the proposed model has the highest score of 97.25% compared to other classifiers on all features.

Table 3
Results of classification using the all textual features

Classifier	Accuracy
MLP	96.67
Logistic regression	45.71
Naïve Bayes (G)	45.58
Nave Bayes(M)	66.39
Naïve Bayes(B)	78.17
Decision Tree	92.28
KNN	95.25
Random Forest	**97.25**
Gradient Boost	95.88
Extra Gradient Boost	94.97
Ada Boost	88.65

In second step, top 14 best features are selected by using Chi-squar features selection technique.All models including proposed model was evaluated using best selected features and results are recorded in Table 4. Experimental results show that proposed model achieved 97.33% accuracy and performed better than individual models on fourteen (14) best features for the task of fake news classification on ISOT dataset.

Table 4
Results of classification using the top 14 features selected by Chi2

Classifier	Accuracy
MLP	92.64
Logistic regression	45.54
Naïve Bayes (G)	65.87
Nave Bayes(M)	66.36
Naïve Bayes(B)	70.17
Decision Tree	93.07
KNN	95.25
Random Forest	**97.33**
Gradient Boost	96.27
Extra Gradient Boost	95.73
Ada Boost	86.70

The above results shows that Random forest attained the highest accuracy score of 97.33%, Gradient Boost obtained the second highest accuracy of 96.27 percent, Extra Gradient Boost accuracy is 95.73 percent, KNN accuracy is 95.25 percent, Decision Tree accuracy is 93.07, MLP accuracy is 92.64 percent and Logistic Regression got the lowest accuracy of 45.54 percent.

In third step, top 14 best features are selected by using the Univariate feature selection technqiue. All models including proposed model are tested using 14 best features and the results are recorded in Table 5. The results demonstrate that the proposed model performed better than individual other models by attaining accuracy score of 97.27%. on best features selected using Univariate features selection technique.

Table 5
Results of classification using the top 14 features selected by Univariate

Classifier	Accuracy
MLP	87.96
Logistic regression	45.90
Naïve Bayes (G)	43.48
Nave Bayes(M)	66.28
Naïve Bayes(B)	78.17
Decision Tree	85.16
KNN	89.99
Random Forest	**97.27**
Gradient Boost	86.69
Extra Gradient Boost	86.40
Ada Boost	84.37

Referring to the results presented in above table, the fourteen best features selected using Univariate feature selection techniques, The proposed model attained highest accuracy of 97.27%, KNN achieved the second highest accuracy of 89.99%, MLP accuracy is 87.96%, Gradient Boost accuracy is 86.69 percent. In this case, Random Forest again remained on top in term of accuracy.

In fourth step, fourteen (14) best features are selected by using the feature importance technique. All models alongwith proposed model are evaluated on top 14 best features and results are stored in Table 6. The results show that proposed model performance is better than other classifier. The accuracy of proposed model is 96.60%.

Table 6
Results of classification using the top 14 features selected by feature importance

Classifier	Accuracy
MLP	96.30
Logistic regression	48.90
Naïve Bayes (G)	48.48
Nave Bayes(M)	70.28
Naïve Bayes(B)	79.17
Decision Tree	92.53
KNN	95.43
Random Forest	**96.60**
Gradient Boost	94.89
Extra Gradient Boost	94.85
Ada Boost	88.78

With reference to the results given in Table 6, the fourteen best features selected using feature importance features selection technique, our proposed model performed better than others by securing highest accuracy of 96.60%. The MLP got the second highest accuracy of 96.30%, KNN accuracy is 95.43%, Gradient Boost accuracy is 94.89 percent.

In fifth step,the best features are selected using information gain method of feature selection. All models including our proposed model are evaluated on top 14 best features selected from ISOT dataset and the result in each case are recorded in Table 7, The proposed model achieved the highest accuracy of 96.42%, MLP achieved the second highest accuracy of 96.02%, KNN accuracy is 95.91%, Gradient Boost accuracy is 94.18 percent, and RF achieved the highest accuracy for individual models and Naive Bayes got lowest accuracy.

Table 7
Results of classification using the top 14 features selected by information gain

Classifier	Accuracy
MLP	96.02
Logistic regression	49.90
Naïve Bayes (G)	56.48
Nave Bayes(M)	65.28
Naïve Bayes(B)	79.17
Decision Tree	92.36
KNN	95.91
Random Forest	**96.42**
Gradient Boost	94.18
Extra Gradient Boost	94.10
Ada Boost	88.50

The results in Fig. 3 show that the accuracy of classifiers moved up and down while using all features, best selected features by Chi2,Univariate, Features importance and information gain.

Table 8
Comparative Results of all classifications over all features and top 14 features selected by Four Features Selection Techniques

Classifier	All features	Chi2	Univariate	Feature Importance	Information gain
MLP	96.67	92.64	87.96	96.30	96.02
Logistic regression	45.71	45.54	45.90	48.90	49.90
Naïve Bayes (G)	45.58	65.87	43.48	48.48	56.48
Nave Bayes(M)	66.39	66.36	66.28	70.28	65.28
Naïve Bayes(B)	78.17	70.17	78.17	79.17	79.17
Decision Tree	92.28	93.07	85.16	92.53	92.36
KNN	95.25	95.25	89.99	95.43	95.91
Random Forest	**97.25**	**97.33**	**97.27**	**96.60**	**96.42**
Gradient Boost	95.88	96.27	86.69	94.89	94.18
Extra Gradient Boost	94.97	95.73	86.40	94.85	94.10
Ada Boost	88.65	86.70	84.37	88.78	88.50

From experimental results shown in Fig. 3 and Table 8, the following conclusion are drawn:

- The accuracy of the proposed model (Random Forest) improved with best features selected by using Chi-square and Univariate. However, the accuracy of proposed model decreased when features selection are perfomed using future importance and information gain features selection techniques.
- The accuracy of boosting technique like XGBoost, GBM and Ada Boost does not improved as a whole on best features .
- Perfomance of Gaussian Naïve Bayesian improved significantly on best features selected by Chi-squar as compared to other features selection techniques.
- Overall, the classification accuracy of the proposed model is superior than all individuals' models as well as other boosting approaches.

5. Conclusion And Future Work

Online fake news detection is a challenging task in the area of text classification. Many attempts have been made by various researchers to address this issue. This study proposed Random Forest as machine learning classifier to classify the news as fake news and real news. For this purpose, twenty three (23) features were extracted from the text of the dataset. Four feature selection techniques like chi-square, Univariate, features importance and information gain, were used to select fourteen (14) best features out of the twenty three (23) extracted features. Proposed model as well as other models were used for

classification of fake news and real news using fourteen (14) best features. The experimental results show that proposed model outperformed all other classifiers in term of better classification accuracy in fake news prediction. In future, Deep Ensembling models may be used for fake news detection.

6. Declarations

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors did not receive support from any organization for the submitted work.

7. References

1. Ahmad I, Yousaf M, Yousaf S, Ahmad MO (2020) Fake news detection using machine learning ensemble methods Complexity 2020
2. Ahmed S, Hinkelmann K, Corradini F Combining machine learning with knowledge engineering to detect fake news in social networks-a survey. In: Proceedings of the AAAI 2019 Spring Symposium, 2019. p 8
3. Al-Ash HS, Putri MF, Mursanto P, Bustamam A Ensemble learning approach on indonesian fake news classification. In: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), 2019. IEEE, pp 1-6
4. Choudhary M, Chouhan SS, Pilli ES, Vipparthi SK (2021) BerConvoNet: A deep learning framework for fake news classification Applied Soft Computing 110:107614
5. De Choudhury M, Morris MR, White RW Seeking and sharing health information online: comparing search engines and social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, 2014. pp 1365-1376
6. Faustini P, Covões T Fake news detection using one-class classification. In: 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 2019. IEEE, pp 592-597
7. Gravanis G, Vakali A, Diamantaras K, Karadais P (2019) Behind the cues: A benchmarking study for fake news detection Expert Systems with Applications 128:201-213
8. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ (2021) An ensemble machine learning approach through effective feature extraction to classify fake news Future Generation Computer Systems 117:47-58
9. Hlaing MMM, Kham NSM Defining news authenticity on social media using machine learning approach. In: 2020 IEEE Conference on Computer Applications (ICCA), 2020. IEEE, pp 1-6

10. Jain A, Shakya A, Khatter H, Gupta AK A smart system for fake news detection using machine learning. In: 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2019. IEEE, pp 1-4
11. Katsaros D, Stavropoulos G, Papakostas D Which machine learning paradigm for fake news detection? In: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2019. IEEE, pp 383-387
12. Khan JY, Khondaker MTI, Afroz S, Uddin G, Iqbal A (2021) A benchmark study of machine learning models for online fake news detection Machine Learning with Applications 4:100032
13. Mahir EM, Akhter S, Huq MR Detecting fake news using machine learning and deep learning algorithms. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC), 2019. IEEE, pp 1-5
14. Mansouri R, Naderan-Tahan M, Rashti MJ A Semi-supervised Learning Method for Fake News Detection in Social Media. In: 2020 28th Iranian Conference on Electrical Engineering (ICEE), 2020. IEEE, pp 1-5
15. Najar F, Zamzami N, Bouguila N Fake news detection using bayesian inference. In: 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), 2019. IEEE, pp 389-394
16. Napoli PM (2018) What if more speech is no longer the solution: First Amendment theory meets fake news and the filter bubble Fed Comm LJ 70:55
17. Okoro E, Abara B, Umagba A, Ajonye A, Isa Z (2018) A hybrid approach to fake news detection on social media Nigerian Journal of Technology 37:454-462
18. Ozbay FA, Alatas B (2020) Fake news detection within online social media using supervised artificial intelligence algorithms Physica A: Statistical Mechanics and its Applications 540:123174
19. Reis JC, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised learning for fake news detection IEEE Intelligent Systems 34:76-81
20. Ruchansky N, Seo S, Liu Y Csi: A hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017. pp 797-806
21. Vogel I, Meghana M Detecting Fake News Spreaders on Twitter from a Multilingual Perspective. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020. IEEE, pp 599-606

Figures

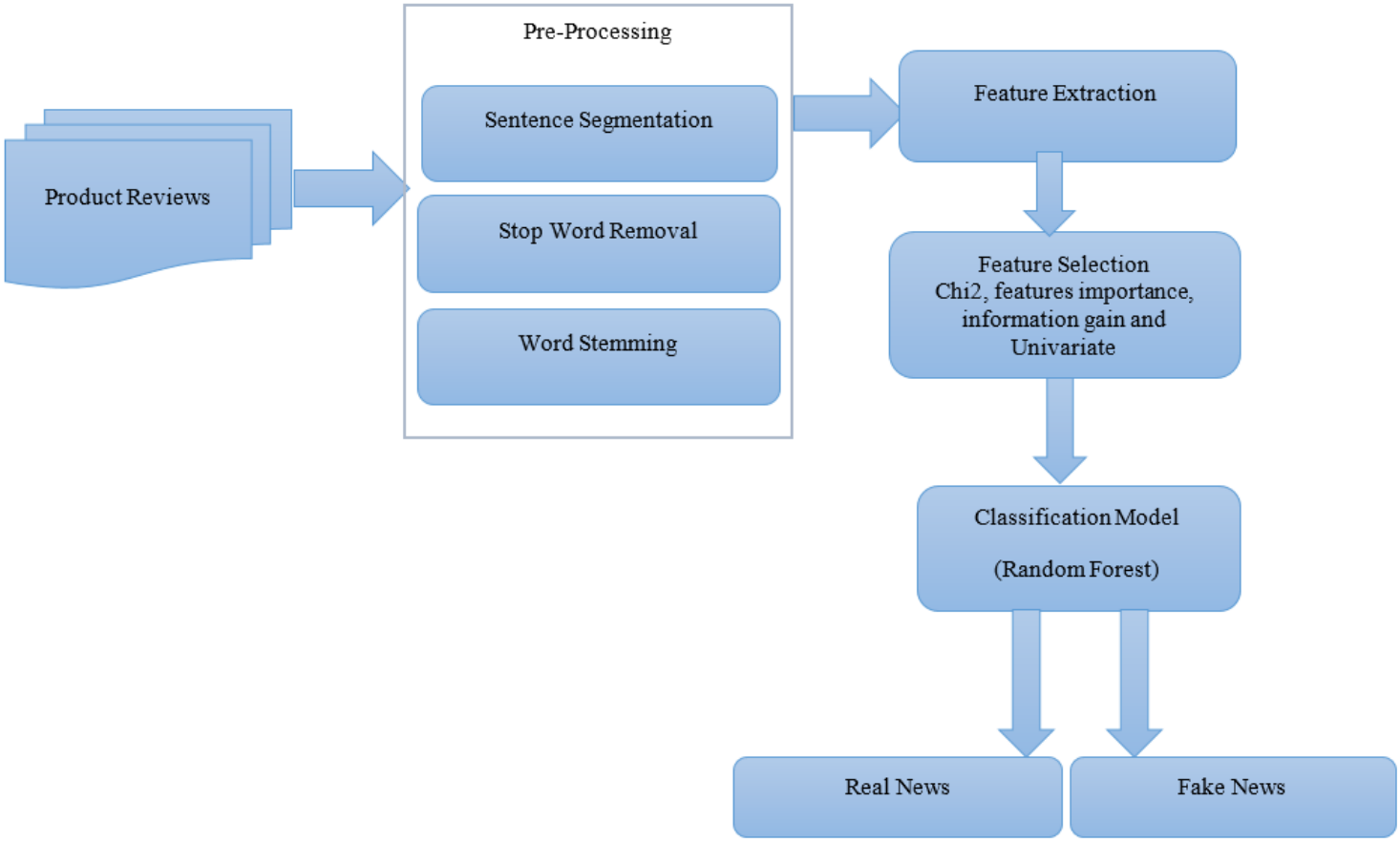

Figure 1

Proposed approach for fake news classification

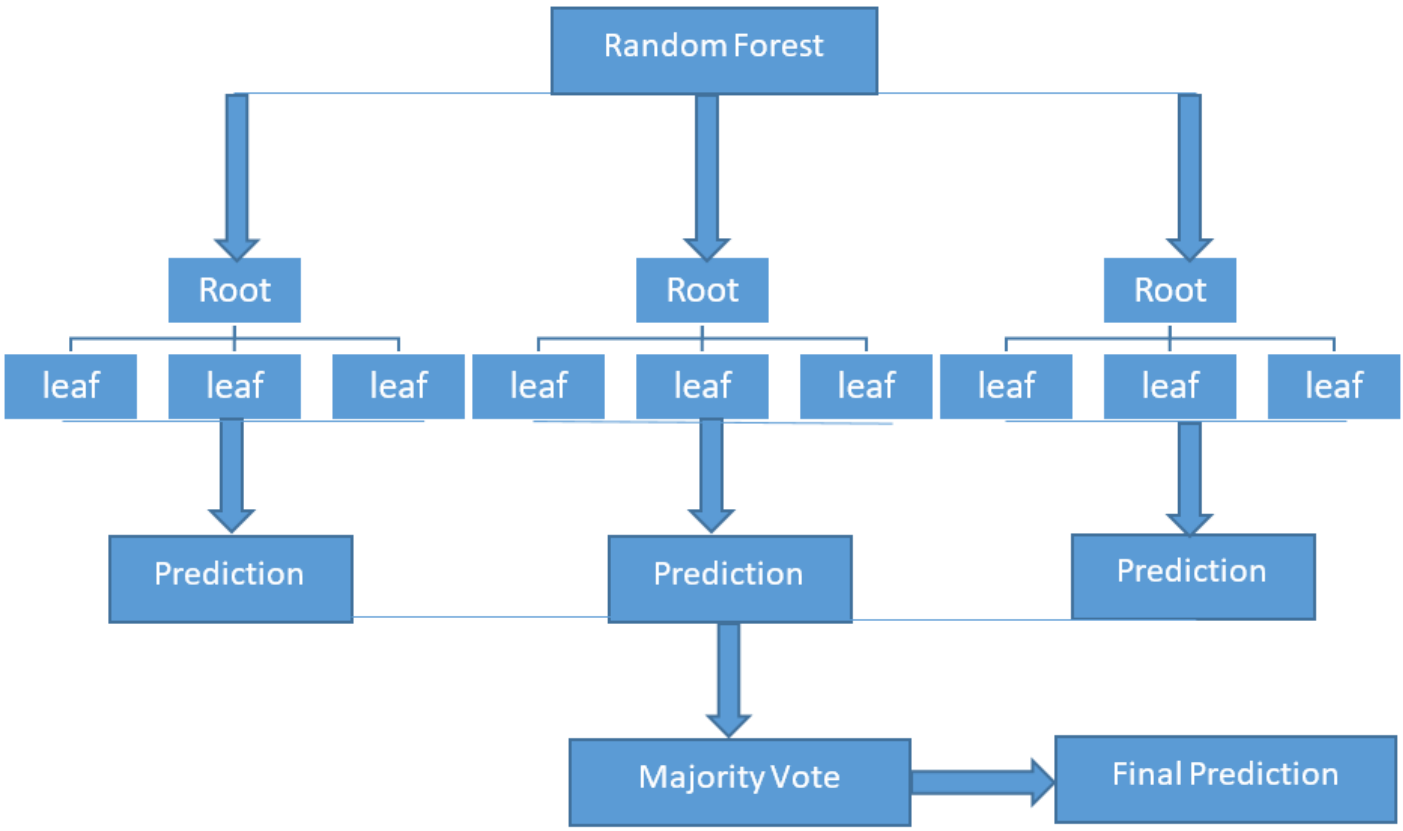

Figure 2

Simple Random Forest classifier structure

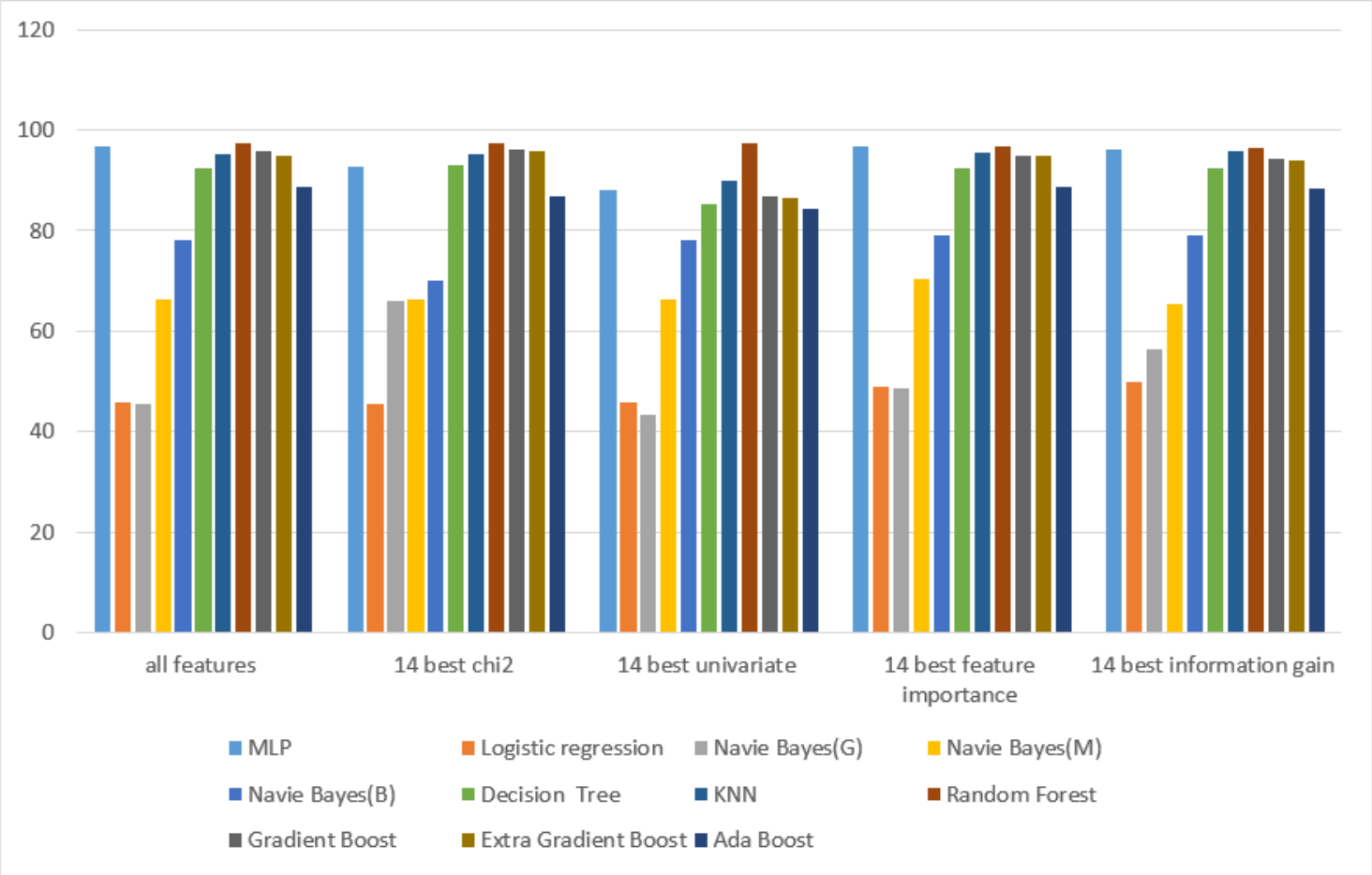

Figure 3

Accuracy of classifiers on all textual features and top 14 features collection obtained for the ISOT News dataset using four feature selection techniques (Chi2, Univariate, feature importance and information gain).