The Review of Socionetwork Strategies (2022) 16:259–289
https://doi.org/10.1007/s12626-022-00127-7

ARTICLE

Enhancing the Predictive Performance of Credibility-Based Fake News Detection Using Ensemble Learning

Amit Neil Ramkissoon[1] · Wayne Goodridge[1]

Received: 23 June 2022 / Accepted: 6 September 2022 / Published online: 17 September 2022

Abstract

Fake news detection continues to be a major problem that affects our society today. Fake news can be classified using a variety of methods. Predicting and detecting fake news has proven to be challenging even for machine learning algorithms. This research employs Legitimacy, a unique ensemble machine learning model to accomplish the task of Credibility-Based Fake News Detection. The Legitimacy ensemble combines the learning potential of a Two-Class Boosted Decision Tree and a Two-Class Neural Network. The ensemble technique follows a pseudo-mixture-of-experts methodology. For the gating model, an instance of Two-Class Logistic Regression is implemented. This study validates Legitimacy using a standard dataset with features relating to the credibility of news publishers to predict fake news. These features are analysed using the ensemble algorithm. The results of these experiments are examined using four evaluation methodologies. The analysis of the results reveals positive performance with the use of the ensemble ML method with an accuracy of 96.9%. This ensemble's performance is compared with the performance of the two base machine learning models of the ensemble. The performance of the ensemble surpasses that of the two base models. The performance of Legitimacy is also analysed as the size of the dataset increases to demonstrate its scalability. Hence, based on our selected dataset, the Legitimacy ensemble model has proven to be most appropriate for Credibility-Based Fake News Detection.

Keywords Credibility-Based Fake News Detection · Decision trees · Ensemble learning · Legitimacy model · Logistic regression · Neural networks

✉ Amit Neil Ramkissoon
amit.ramkissoon@my.uwi.edu

Wayne Goodridge
wayne.goodridge@sta.uwi.edu

[1] Department of Computing & Information Technology, The University of the West Indies at St Augustine, St Augustine, Trinidad and Tobago

1 Introduction

Fake news is a classification of information shared in today's world that has become significantly popular. Fake news can be defined as any news article that is intentionally and verifiably false [26]. One example of the popularity of fake news is provided by [23] wherein 171,365 stories from 60 fake news websites for the period between 2014 and 2016 were discovered. As stated in [27] allowing fake news to spread especially via social media and internet-enabled platforms erodes the long-standing efforts that have been made and achieved against the spread of misinformation. Also, allowing this spread of intentionally false news allows for mistruths, misinformation, and spurious conspiracies to establish a living ecosystem. Therefore, it is critical to detect fake news on social media to mitigate these negative effects to benefit the public and the news ecosystem.

Predicting and detecting fake news is a challenge. Many consumers of online news especially via social media and internet-enabled platforms do not cross-reference news posted to such Websites and as such are unable to verify the credibility of the news. Hence, automated credibility validation of news is a necessity [30]. According to [27], detecting fake news on social media is quite challenging since fake news is written to intentionally mislead users, using different styles while mimicking the real news. Fake News detection, therefore, has gathered momentum and is now receiving great attention from the research community.

A major methodology used for the prediction of fake news is the analysis of the features of fake news to determine if any relationships exist amongst these features. According to [35], fake news detection is subdivided into four categories based on the perspective of the detection strategy. These perspectives as stated in [35] are (i) knowledge-based, which focuses on the information in the published news and fact checks this information to determine whether or not it is fake; (ii) style-based, which focuses on how the published news is written and checks for subtle similarities with the styles of both genuine and fake news; (iii) propagation-based, which focuses on the travel pattern of the published news and how it spreads, and (iv) credibility-based, which investigates the validity of the published news based upon the credibility of the publisher and spreaders.

As stated above in (iv), credibility-based fake news detection focuses on detection techniques based upon the established reputation of the news publisher, if the news publisher is seen as a person of ill-repute, then the news cannot be trusted and vice versa. According to [35], when attempting to detect fake news based upon the credibility of the publisher, the detection involves data surrounding the online social behaviour of the publisher as well as the news-related information. For example, a message published by an unreliable publisher and forwarded by unreliable network clients is less likely to be trustworthy and more likely to be fake news than news posted by authoritative and credible users.

As shown in [10], the author/editor/publisher of the news document is highly related to the news content and context. News context is the social engagement of the news article and its consumption on social media platforms. This social engagement represents the news propagation over time and the group of users that engaged with this news document. Hence, social-based features such as the number of followers, friends count, registration age, number of authored posts/tweets, related social groups, demographic information, user stance, average and credibility scores, for example, can be extracted for users, groups, and postings.

To accomplish the task of credibility-based fake news detection, many features of the users have to be investigated to understand how they indicate the credibility of the news. As illustrated in [26], fake news pieces are likely to be created and spread by non-human accounts, such as social bots or cyborgs. Thus, capturing users' profiles and user-based features of those who have interacted with the news on social media can provide useful information for fake news detection.

This research utilizes credibility-based features with an ensemble learning methodology to predict whether the news is fake or not based on the publisher's credibility. Machine learning (ML) is used to recognise patterns in the data which lead to a prediction. One of the scientific fields and probably the most successful, which specializes in tackling any kind of learning task, is the ML field [18]. The two major categories of tasks the ML field usually deals with are classification and regression; however, only classification will be considered in the context of this work, because classification finds patterns between the features of fake news and the legitimacy of the fake news.

This paper expands on the use of Legitimacy, a unique ensemble learning model for the task of Credibility-based Fake News Detection. This model consists of two underlying techniques; namely, a Two-Class Boosted Decision Tree and a Two-Class Neural Network. This research attempts to enhance the predictive performance of credibility-based fake news detection by utilising an ensemble learning prediction model that has previously been introduced in [21]. This research attempts to answer the following question, "Can the predictive performance of credibility-based fake news detection be enhanced with the use of an ensemble learning model for prediction"?

According to [8], decision trees constitute a class of predictive models that score well on efficiency and comprehensibility, mainly because of their simplicity. They use a process where the data is recursively split into smaller and purer subsets by repeatedly applying a greedy search through the space of possible decision trees' branches and choosing optimal splits based upon a splitting criterion. This process starts in the root node, which is a node without parent nodes, and iteratively determines optimal splitting criteria that divide the data over two child nodes. This process terminates, when no further splits are desirable or possible, with a set of nodes without child nodes called terminal nodes or leaves.

To deal with the problems of decision trees, 'boosting' is applied. According to [9], boosting is a general approach that can be applied to many base learners for regression or classification. Boosting converts a learning algorithm for a base class of models with weak predictive power, such as decision trees, into a learning algorithm for a class of models with stronger predictive power, such as a weighted majority vote over base models in the case of classification, or a linear combination of base models in the case of regression. In boosting, base learners are trained iteratively to increase emphasis on observations modelled poorly by the existing collection of base learners. Different boosting algorithms differ in how they quantify misclassification and in their selection of settings for the next iteration.

Alongside the Boosted Decision Tree model, a Two-Class Neural Network is implemented. According to [15], a Neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units.

Neural networks involve long training times and are, therefore, more suitable for applications where this is feasible. They require several parameters that are typically best determined empirically such as the network topology or 'structure.' Neural networks have been criticized for their poor interpretability. For example, humans find it difficult to interpret the symbolic meaning behind the learned weights and 'hidden units' in the network. These features initially made neural networks less desirable for data mining.

This work combines these two models to build an ensemble learning model. Ensemble Learning as defined in [13] is a technique that blends the predictions of several machine learning-based algorithms to make more accurate predictions. In other words, in an ensemble approach, multiple learning models are trained to create a single powerful predictive model.

According to [3], ensemble learners tend to have higher accuracies, as more than one model is trained using a particular technique to reduce the overall error rate and improve the performance of the model. The intuition behind ensemble modelling is synonymous with the intuitions already used in our daily life such as when requesting the opinions of multiple experts before making a particular decision to minimize the chance of a bad decision or an undesirable outcome. For example, a classification algorithm can be trained on a particular dataset with a unique set of parameters that can produce a decision boundary which fits the data to some extent. The outcome of that particular algorithm depends not only on the parameters that were provided to train the model, but also on the type of training data. If the training data contain less variance or uniform data, then the model might overfit and produce biased results over unseen data. Therefore, approaches like cross-validation are used to minimize the risk of overfitting. Several models can be trained on different sets of parameters to create multiple decision boundaries on randomly chosen data points as training data. Hence, using ensemble learning techniques, these problems can be addressed and mitigated by training multiple algorithms, and their results can be combined for a near-optimum outcome.

The objectives of this paper are as follows:

1. To present a review of the current models used for fake news detection and their limitations.
2. To improve the performance of credibility-based fake news detection by utilising an ensemble learning prediction model.
3. To analyse the Legitimacy ensemble that uses credibility to detect fake news
4. To present and evaluate the initial results of the model for credibility-based fake news detection

The research contributions of this paper are as follows:

1. Analysis of a novel ensemble learning model for credibility-based fake news detection
2. Evaluation of the performance of the proposed method using experimentation
3. Comparison of the performance of the proposed ensemble method against supervised methods

The remainder of this paper is structured as follows. Section 2 presents related work in this field, while Sect. 3 presents an understanding of the machine learning algorithm. Experiments are conducted in Sect. 4 using the algorithm and the results of these are discussed in Sect. 5. Section 6 presents a comparative analysis of the ensemble and its composite models, while Sect. 7 presents a scalability analysis of the proposed Legitimacy model. Section 8 discusses the limitations of the Legitimacy model and Sect. 9 concludes this paper.

2 Literature Review

Prior research in fake news detection and machine learning has been conducted. As stated in [27], Model-oriented fake news research opens the door to building more effective and practical models for fake news detection. Most previously mentioned approaches focus on extracting various features, incorporating these features into supervised classification models, such as naive Bayes, decision trees, logistic regression, k nearest neighbour (KNN), and support vector machines (SVM), and then selecting the classifier that performs the best. More research can be done to build more complex and effective models and to better utilise extracted features, such as aggregation methods, probabilistic methods, ensemble methods, or projection methods.

According to [14], fake news, particularly with the speed and reach of unverified/false information dissemination, is a troubling trend with potential political and societal consequences, as evidenced in the 2016 United States presidential election, the ongoing COVID-19 pandemic, and ongoing protests around the world. To mitigate such threats, a broad range of approaches have been designed to detect and mitigate online fake news. In [14], the authors systematically review existing fake news mitigation and detection approaches and identify many challenges and potential research opportunities (e.g., the importance of a data-sharing platform that can also be used to facilitate machine/deep learning).

According to [6], an overview of the various models in detecting fake news such as Machine learning, Natural Language Processing, Crowdsourced techniques, Expert fact-checker, as well as Hybrid Expert-Machine are introduced. The aforementioned research also examined different types of fake news, which is an essential criterion for detection. The findings show that detecting fake news is a challenging but workable task. The techniques which combine people and machines bring very satisfactory results. Furthermore, as stated in [6], Early Machine Learning methods in detecting fake news assume fake news is created intentionally for political and financial benefit, so that such fake news often has an opinionated and enticing headline. Therefore, the extraction of the textual and linguistic feature is necessary for ML. The Naive Bayes classifier and classified linguistic features such as lexical features, including word count and level, as well as syntactic differences, which involve sentence level characterization, have been used. This work, however, has ignored the significant role that the credibility of the publisher plays in detecting fake news.

In [30], an overall performance analysis of different approaches on three different datasets was presented. It was then shown that Naive Bayes with n-gram can attain analogous results to neural network-based models on a dataset with less than 100 k news articles. The performance of Long Short-Term Memory (LSTM) based models greatly depends on the length of the dataset as well as information given in a news article. With adequate information provided in a news article, LSTM-based models have a higher probability to overcome overfitting. Moreover, advanced models like Convolutional neural networks (CNNs) LSTM (C-LSTM), Conv HAN and character level C-LSTM have shown high promise which demands further attention on these models in fake news detection. Finally, a topic-based analysis that exposes the difficulty to detect political, health and research-related deceptive news was performed. This work ignored the use of ensemble learning in fake news detection and only focused on supervised learning approaches.

As illustrated in [12], several models are utilised to detect fake news online. Four different models to detect fake news have been utilised mainly. The Naïve Bayes, Support Vector Machine (SVM), Logistic Regression and the Multi-layer Perceptron methodologies have been utilised to detect fake news. An SVM was applied to a dataset consisting of 12,600 truthful articles and 12,600 fake articles. First, the dataset was pre-processed using stop word removal and stemming features were extracted using term frequency and term frequency–inverse document frequency (TF-IDF) and a feature matrix was formed from the documents. This was then passed through a classifier. The classifier consists of six different machine learning algorithms, Stochastic Gradient Descent (SGD), SVM, Linear Support Vector Machine (LSVM), KNN and Decision Trees (DT). The highest accuracy was obtained when using unigrams features and linear SVM, which gave an accuracy of 92%. This work is similar to that conducted in [22]; however, like previous authors, it ignores the strength of ensemble learning prediction in fake news detection.

As reported in [34], supervised machine learning algorithms like DT, Random Forest, SVM, Logistic Regression and KNN have been used extensively in previous works for online hoaxes, frauds, and deceptive information classification.

According to [19], the revolution in social media has propelled the online community to take advantage of online reviews not only for posting feedback about products and services, but also to assist individuals in analysing a user's feedback to make purchasing decisions, and for companies to improve the quality of manufactured goods. However, the propagation of fake reviews has become an alarming issue, as it deceives online users while purchasing and promotes or demotes the reputation of competing brands. In [19], the authors propose a supervised learning-based technique for the detection of fake reviews from online textual content. Their study employs machine learning classifiers to bifurcate fake and genuine reviews. Experimental results are evaluated against different evaluation measures and the performance of their proposed system is compared with baseline works. Their research only investigates the textual features of the news and utilises only supervised learning for prediction rather than ensemble learning.

As stated in [3], ensemble learning has been used previously in Fake News Detection. According to [3], in current fake news corpora, multiple instances have occurred where both supervised and unsupervised learning algorithms are used to classify text. However, the authors explain that most of the literature focuses on specific datasets or domains, most prominently the domain of politics. Therefore, the algorithm trained works best on a particular type of article's domain and does not achieve optimal results when exposed to articles from other domains. Since articles from different domains have a unique textual structure, it is difficult to train a generic algorithm that works best on all particular news domains.

Their study explores different textual properties that could be used to distinguish fake content from real. Using those properties, they train a combination of different machine learning algorithms using various ensemble methods that are not thoroughly explored in the current literature. The ensemble learners have proven to be useful in a wide variety of applications, as the learning models tend to reduce the error rate using techniques such as bagging and boosting. These techniques facilitate the training of different machine learning algorithms effectively and efficiently. The authors also conducted extensive experiments on 4 real-world publicly available datasets. The results validate the improved performance of their proposed technique using the 4 commonly used performance metrics (namely, accuracy, precision, recall, and F1 score). In their paper, they propose a solution to the fake news detection problem using the machine learning ensemble approach; however, they focus only on the textual analysis and ignore the importance of the publisher's credibility.

According to [13], there are numerous channels available such as social media, blogs, websites, for example, through which people can easily access the news. The availability of these platforms has made the dissemination of fake news easier. Anyone using these platforms can create and share fake news content based on personal or professional motives. To address the issue of detecting fake news, numerous studies based on supervised and unsupervised learning methods have been proposed. However, all those studies suffer from a certain limitation of poor accuracy. The reason for poor accuracy can be attributed to several reasons, such as the poor selection of features, inefficient tuning of parameters, and imbalanced datasets, for example. In their article, the authors have proposed an ensemble classification model for the detection of fake news that has achieved a better accuracy compared to the state-of-the-art. The proposed model extracts key features from the fake news datasets, and the extracted features are then classified using the ensemble model consisting of three popular machine learning models, namely, the Decision Tree, Random Forest, and the Extra Tree Classifier. They achieved a training and testing accuracy of 99.8% and 44.15%, respectively, on the ISOT dataset. For the Liar dataset, they achieved the training and testing accuracy of 100%. Their model does not use features that cover the entire spread of those related to fake news and it ignores the combination of Boosted Decision Trees and Neural Networks.

According to [17] in their research paper, first, they investigated the existing models for fake news detection using content and context-based information. After an initial investigation, the authors performed extensive experiments using a multi-class dataset (FNC-based fake news dataset) and employed different machine learning algorithms. In their exploration, they have found that among the different machine learning algorithms used, Gradient Boosting with optimized parameters performs the best for a multi-class fake news dataset. In the existing research, benchmark results are available based on two classes in the dataset, namely, classifying news as fake or real. Work on multi-class prediction is limited. There is a huge scope for improvement in multi-class fake news detection. Their research is an attempt to improve on the existing fake news classification using a multi-class dataset with the motivation that it can be helpful for future researchers working in this area. However, their research ignores the use of any other model than that of gradient boosting.

According to [24], most of the existing studies on fake news detection are based on the classical supervised model. In recent times, there has been an interest in developing deep learning-based fake news detection systems, but these are mostly concerned with binary classification. In their paper, they attempt to develop an ensemble-based architecture for fake news detection. The individual models are based on Convolutional Neural Networks (CNNs) and Bi-directional Long Short-Term Memory (LSTM). The representations obtained from these two models are fed into a Multi-layer Perceptron (MLP) for multi-class classification. This ensemble model employs the use of computationally heavy supervised models to detect fake news, which may make it inappropriate for energy-constrained devices.

Table 1 below presents a comparison of the related work by listing their key characteristics and their associated limitations. In the end, the proposed work of this research is also presented.

3 Detection Model

This research analyses the performance of an ensemble learning model for fake news detection. Legitimacy is based upon two underlying models. Legitimacy merges a Two-Class Boosted Decision Tree and a Two-Class Neural Network to form a single, powerful classifier to categorise news as either fake or genuine. Legitimacy is built based upon a pseudo-mixture-of-experts ensembling framework, which allows for multiple classification models to be used and trained on different subsets of the feature set. The classification results from these individual models which are combined using a gating model. Two-Class Logistic Regression is employed as the gating model of Legitimacy and uses the combined classification results of the individual models to make the final decision.

Legitimacy is based upon models proposed by Microsoft Azure Machine Learning Studio (classic) (AzureML). AzureML is a collaborative, drag-and-drop tool you can use to build, test, and deploy predictive analytics solutions on your data [16]. It publishes models as Web services that can easily be consumed by custom apps or business intelligence tools. AzureML is where data science, predictive analytics, cloud resources, and data meet.

Table 1 Comparison of related works

Related work	Key characteristics	Limitations
[6]	Investigates ML methods alongside fake news features. Determines that Naïve Bayes and linguistic features are best suited for fake news detection	Their work, however, has ignored the significant role that the credibility of the publisher plays in detecting fake news
[30]	Investigates ML methods for fake news detection. Determines that Naïve Bayes is best for fake news detection	Their work ignores the use of ensemble learning in fake news detection and only focuses on supervised learning approaches
[12]	Investigates ML methods for fake news detection. Determines the highest accuracy was obtained when using unigrams features and linear SVM	This work is similar to that conducted in [22], however, like previous work, it ignores the strength of ensemble learning prediction in fake news detection
[19]	In their work, they propose a supervised learning-based technique for the detection of fake reviews from online textual content	Their work only investigates the textual features of the news and utilises only supervised learning for prediction rather than ensemble learning
[3]	Their study explores different textual properties that can be used to distinguish fake content from real by training different models with the fake news features	In their paper, they propose a solution to the fake news detection problem using the machine learning ensemble approach, however, they focus only on the textual analysis and ignore the importance of the publisher's credibility
[13]	In their article, they proposed an ensemble classification model for the detection of fake news that has achieved a better accuracy compared to the state-of-the-art. The proposed model extracts key features from the fake news datasets, and the extracted features are then classified using the ensemble model consisting of three popular machine learning models namely, Decision Tree, Random Forest, and Extra Tree Classifier	Their model does not use features that cover the entire spread of those related to fake news as well and the model ignores the combination of Boosted Decision Trees and Neural Networks
[17]	In their exploration, they found that among the different machine learning algorithms used, Gradient Boosting with optimized parameters performs the best for a multi-class fake news dataset	Their research is an attempt to improve the existing fake news classification using a multi-class dataset with the motivation that it can be helpful for future researchers working in this area. However, their research ignores the use of any other model than that of gradient boosting
[24]	In their paper, they attempt to develop an ensemble-based architecture for fake news detection. The individual models are based on Convolutional Neural Networks (CNN) and Bi-directional Long Short-Term Memory (LSTM)	This ensemble model employs the use of computationally heavy supervised models to detect fake news which may make it inappropriate for energy-constrained devices

Table 1 (continued)

Related work	Key characteristics	Limitations
Proposed Work	Legitimacy is a unique ensemble learning model for the task of Credibility-based Fake News Detection. This pseudo-mixture-of-experts model consists of two underlying techniques namely, a Two-Class Boosted Decision Tree and a Two-Class Neural Network using Two-Class Logistic Regression as a gating model. This research attempts to improve the performance of credibility-based fake news detection by utilising an ensemble learning prediction model	Legitimacy is based on credibility and content-based fake news detection. Other areas of its application are yet to be explored

The ensemble model consists of the following classification models as described in [16] and they are chosen based on the experiments conducted by [22].

Two-Class Boosted Decision Tree (BDT) The Two-Class Boosted Decision Tree model has been proposed by Microsoft Azure Machine Learning Studio (classic). This paper utilises the Two-Class Boosted Decision Tree based on the results stated in [22] and further experimentation. According to [22], from the experiments performed, and the results obtained, the Two-Class Boosted Decision Tree performed the best. Hence, it can be concluded that based on our selected dataset the Two-Class Boosted Decision Tree is the best method suited for detecting and predicting Credibility-Based Fake News.

As described in [16], a boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction. Generally, when properly configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks. However, they are also one of the more memory-intensive learners, and the current implementation holds everything in memory. Therefore, a boosted decision tree model might not be able to process the very large datasets that some linear learners can handle.

Boosting can take the form of two algorithms. According to [9] in the first method, the AdaBoost algorithm constructs an ensemble by focusing on instances that were previously misclassified. The level of focus is determined by assigning weights to the instances in the training set. The same weight is assigned to all the instances in the training set during the first iteration. With the rise in the number of iterations, a rise in the weights of misclassified instances occurs. On the other hand, the weights of correctly classified instances are gradually reduced. Additionally, when making a prediction using the generated ensemble, weights are also assigned to the individual base learners by considering their overall predictive performance. Ensemble construction using AdaBoost has been adapted from the work of [28]. Despite the focus of boosting primarily being on bias reduction, slight variance reduction can be achieved by reweighting, as is the case in AdaBoost. Variance reduction occurs because of the construction of models iteratively on randomly sampled, but reweighted, training instances. The reweighting scheme controls the amount of variance reduction. For classification trees, low bias and low variance can be achieved since decision trees are a low bias and high variance technique. In their work, the ensemble constructed using AdaBoost with decision stumps as weak learners and serves as a baseline for comparing prediction metrics with that of scalable GBDT systems.

The gradient-descent-based formulation of boosting methods and the corresponding models termed gradient boosting machines (GBMs) is the second method as described by [9]. GBMs construct base learners iteratively by reweighting observations that were misclassified. However, GBMs differ from AdaBoost in that GBMs determine the weights by operating on the negative partial derivatives of the loss function at each training observation. These partial derivatives are also called pseudo-residuals and an ensemble is grown iteratively using these pseudo-residuals. Consequently, the feature space is partitioned grouping similar pseudo-residuals together. While GBMs can be efficient for relatively small datasets, for much larger datasets, scalable versions are needed. XGBoost, LightGBM and CatBoost are recently developed tree-based scalable versions of GBMs designed to address this requirement. They distinguish the originally formulated GBMs that use decision trees as base learners from the scalable versions by labelling them as gradient-boosted decision classifiers (GBDCs). In their work, GBDCs serve as a baseline for comparison of performance metrics with that of scalable GBDT systems, namely, XGBoost, LightGBM and CatBoost. The Two-Class Boosted Decision Tree implements the decision tree algorithm and boosts the tree utilising the GBM methodology.

The GBM equation can be seen in (1):

$$\left(\rho_t, \theta_t\right) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} -g_t\left(x_i\right) + \rho h\left(x_i, \theta\right). \tag{1}$$

The exact form of the derived algorithm with all the corresponding formulas will heavily depend on the design choices of $\psi(y,f)$ and $h(x, \theta)$.

Two-Class Neural Network A Two-Class Neural Network is chosen as the second model of this ensemble based on experiments conducted in [22] and further experimentation. According to [16], a neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes.

Between the input and output layers, multiple hidden layers can be inserted. Most predictive tasks can be accomplished easily with only one or a few hidden layers. However, recent research has shown that deep neural networks (DNN) with many layers can be very effective in complex tasks such as image or speech recognition. The successive layers are used to model increasing levels of semantic depth.

The relationship between inputs and outputs is learned from training the neural network on the input data. The direction of the graph proceeds from the inputs through the hidden layer and to the output layer. All nodes in a layer are connected by the weighted edges to nodes in the next layer.

To compute the output of the network for a particular input, a value is calculated at each node in the hidden layers and the output layer. The value is set by calculating the weighted sum of the values of the nodes from the previous layer. An activation function is then applied to that weighted sum. The softmax function can be defined as:

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{c} e^{x_j}}. \tag{2}$$

According to [2], the ANN is found to be a very novel and useful model applied to problem-solving and machine learning. An ANN is an information management model that is like the biological nervous system function of the human's brain. Recently, research interest in brain functionality has rapidly increased globally. According to [2], an ANN can be a comparable machine produced to function the same way the human brain performs a given task of interest. For example, "the human brain is big and highly efficient. The human brain is like an information-processing machine that has a variety of complex signal computing operations," that can be easily coordinated to perform a task. The main element of this brain is the unique design of its information-processing capability. It constitutes many complex and interconnected "neurons" in the form of elements working together to solve specific problems on daily basis. A typical example of a neural network function is the human brain that is connected to send and receive signals for human action.

As stated in [2], neural network (NN) layers are independent of one another; that is, a specific layer can have an arbitrary number of nodes. This arbitrary number of nodes is called a bias node. The bias nodes are always set as equal to one. Analogously, the bias nodes are like the offset in a linear regression given as; $y = ax + b$, where a is the coefficient of independent x and then b is called a slope. A bias's major function is to provide a node with a constant value that is trainable, in addition to the normal inputs received by the network node. Importantly, a bias value enables one to move the activation function either to the right or the left, which can be analytical for ANN training success. When the NN is used as a classifier, the input and the output nodes will match input features and output classes. However, when the NN is used as a function approximation, it has an input and an output node. However, the number of designed hidden nodes is essentially greater than those of input nodes.

Mixture-of-Experts In [31], the original ME regression and classification models are described. In the ME architecture, a set of experts and a gate cooperate to solve a nonlinear supervised learning problem by dividing the input space into a nested set of regions used for classification. The gate makes a soft split of the whole input space, and the experts learn the simple parameterized surfaces in these partitions of the regions. The parameters of these surfaces in both the gate and the experts can be learned using the EM algorithm.

Two-Class Logistic Regression Logistic regression is a well-known statistical technique that is used for modelling many kinds of problems. It is used to predict the probability of an outcome and is especially popular for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function. This algorithm is a supervised learning method; therefore, a dataset must be presented that already contains the outcomes to train the model.

According to [20], linear models are composed of one or multiple independent variables that describe a relationship to a dependent response variable. Mapping qualitative or quantitative input features to a target variable that is to be predicted, such as financial, biological, or sociological data is known as supervised learning in machine learning terminology if the labels are known. One of the most utilized linear statistical models for discriminant analysis is logistic regression.

Simplicity and interoperability of logistic regression can occasionally lead to outperforming other sophisticated nonlinear models such as ensemble learners or support vector machines. However, in the event, the response variable is drawn from a small sample size, then logistic regression models become insufficient and perform poorly for binary responses. Any learning algorithm could be applied to modelling binary classification data types; however, the focal point of this work is to examine one linear model, logistic regression.

In the case of logistic regression, the response variable is quantitative. For logistic regression, the response variable is the log of the odds of being classified in the ith group of a binary or multi-class response. Logistic regression makes several assumptions such as independence, responses (logits) at every level of a subpopulation of the explanatory variable which are normally distributed, and constant variance between the responses and all values of the explanatory variable. Intuitively, a transformation to the response variable is applied to yield a continuous probability distribution over the output classes bounded between 0 and 1. This transformation is called using the "logistic" or "sigmoid" function where 'z' corresponds to log odds divided by the logit. The parameter estimates inform whether there is an increase or decrease in the predicted log odds of the response variable that would be predicted by one unit increase or decrease in one of the explanatory variables (e.g., $\times 1$), while holding all other explanatory variables constant.

By combining the above individual learning models, the Legitimacy ensemble model presented in this paper is built. Legitimacy combines the functionality of a Two-Class Boosted Decision Tree and a Two-Class Neural Network. The ensemble model is designed as seen in Fig. 1.

From the dataset, the features are separated into two distinct subgroups namely demographic features and social behaviour features. The ensemble model is built based on these two separable groups, to maximise the performance of each individual model based on one of the subgroups. Once the data have been separated the normal data processing rules follow from here where the data are cleaned and normalized. The data is split using a 65–35% split for training and testing with 65% used for training and 35% used for testing. The Two- Class Boosted Decision Tree is applied to the social behaviour data features, given its excellent performance with this subgroup.

The Two-Class Neural Network model is applied to the demographic features, because this model shows its excellent behaviour with this group of features. Following the rules of machine learning and ensemble learning each individual model is trained using the subgroup of features chosen. Both models are then scored and evaluated individually. This scored dataset is scored with the testing dataset. Each scored dataset is combined to form a larger scored dataset and this amalgamated scored dataset is used to train the logistic regression model. The Two-Class Logistic Regression uses the output of both models and then trains itself with this data. This model is chosen based on experimentation with different models as the gating model. Two-Class Logistic Regression was found to have the best performance as the gating model and hence, chosen as the gating model.

As stated in [7], Logistic Regression can be defined using the formula found in (3):

Let Y denote the binary response variable of interest and X_1, …, X_p the random variables considered as explaining variables, termed features in this paper. The logistic regression model links the conditional probability $P(Y = 1|X_1, ..., X_p)\, to\, X_1, ..., X_p$ through

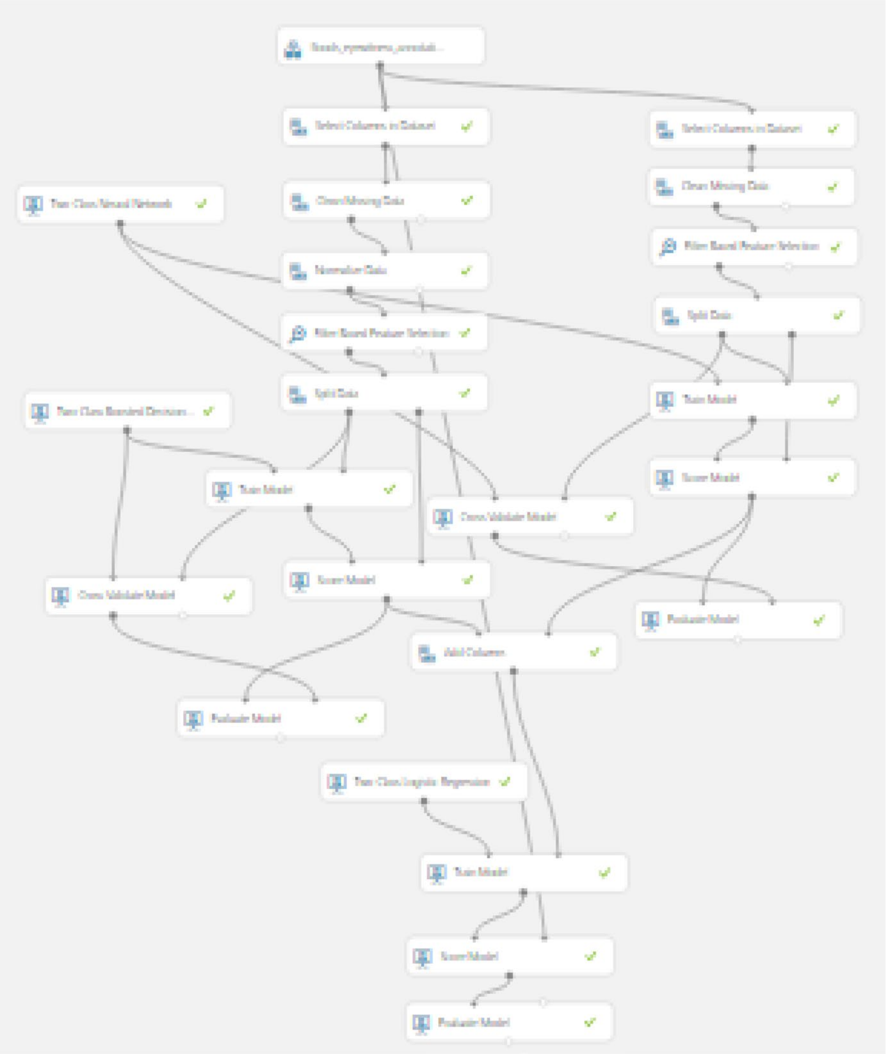


Fig. 1 Experimental setup

$$P(Y = 1|X_1, \ldots, X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}, \tag{3}$$

where β_0, β_0,…,β_p are regression coefficients, which are estimated by maximum-likelihood from the considered dataset. The probability that $Y = 1$ for a new instance is then estimated by replacing the β's by their estimated counterparts and the X's by their realizations for the considered new instance in (3).

The Logistic Regression model works as the gating model in this pseudo-mixture-of-experts model. The Logistic Regression uses the merged dataset and performs a weighted combination to train itself. The scored features of the merged dataset will be the list of X's as seen above. The class variable, which in this case is whether the news is true or fake, is the predicted Y, which is what this work attempts to achieve. The Logistic Regression model chooses a percentage of each data feature to use in each situation and applies this to training the model. The model is scored and evaluated. The scored result of this Logistic Regression is then utilised as the final predicted value of the model. From the analysis done, the computational complexity was seen to be $O(n^2)$.

4 Experimental Design

The focus of this research uses a methodology that is assessed using AzureML.

The following are the test environment parameters:

Dataset The dataset used as proposed by [32] is described as a resource comprising 14,000 labelled tweets collected during several natural disasters including hurricanes, earthquakes, floods, and forest fires. The tweets were annotated following an eyewitness taxonomy defined in their paper. Their paper aims to address this research gap by designing an eyewitness reports taxonomy focusing on the needs of disaster response agencies during natural disasters. Moreover, their research explores several types of features associated with tweets to train machine learning models for the automatic classification of tweets. For this purpose, distinct characteristics (mainly language-based) from tweets posted by eyewitnesses are manually learned. These characteristics are used as independent (domain-expert) features together with content-based features from tweets to train several machine learning models. The features of the tweets include the text, eyewitness label, source, date/time, language, listed count, location, statuses count, followers count, favourites count, time zone, user language, friends count and screen name. Since creating a balanced labelled dataset from social media labelled data is a challenging task, a state-of-the-art class balancing technique is implemented by the authors. They demonstrate that a model trained on a balanced dataset can achieve even higher results.

The experiment is conducted using the above machine learning method, using the same steps and the dataset to enhance objectivity. For the algorithm, the experimental setup for the experiment is like most machine learning experiments. The entire dataset of 14 features is separated into two subsets, the social behaviour subset consisting of 8 features namely, text, eyewitness label, source, listed count, statuses count, followers count, favourites count and friends count. The demographic subset consisting of the remaining 6 features namely, date/time, language, time zone, user language, location and screen name. This division is selected after a thorough sensitivity analysis was performed and the 8–6 split was shown to produce the best results. The data from the dataset is cleaned for missing data by replacing the missing data with the mode of that data feature. The mode is chosen as the replacement strategy as the data is a mix of both numeric and categorical data and as such methods like the mean and median are invalid and only the mode is valid. The data are then split using a proportion of 65% for training data and 35% for testing. The training data are then used with the selected machine learning algorithm. These results are then scored and evaluated.

5 Results and Observations

For the ensemble model proposed in this research, the experimental results are shown below. The results are evaluated based on the Accuracy, Precision, Recall, $F1$-score, and AUC values as well as three types of graphs.

A. ROC Curve

As illustrated in Fig. 2, the results are first analysed based on their respective receiver operating characteristics (ROC) curves. A ROC curve is a technique for visualizing, organising and selecting classifiers based on their performance [11]. ROC curves have long been used in signal detection theory to depict the trade-off between hit rates and the false alarm rates of classifiers.

B. Precision/Recall Curve

The second method used to analyse the performance of the machine learning algorithm is the Precision/Recall Curve as seen in Fig. 3. Precision is a ratio of

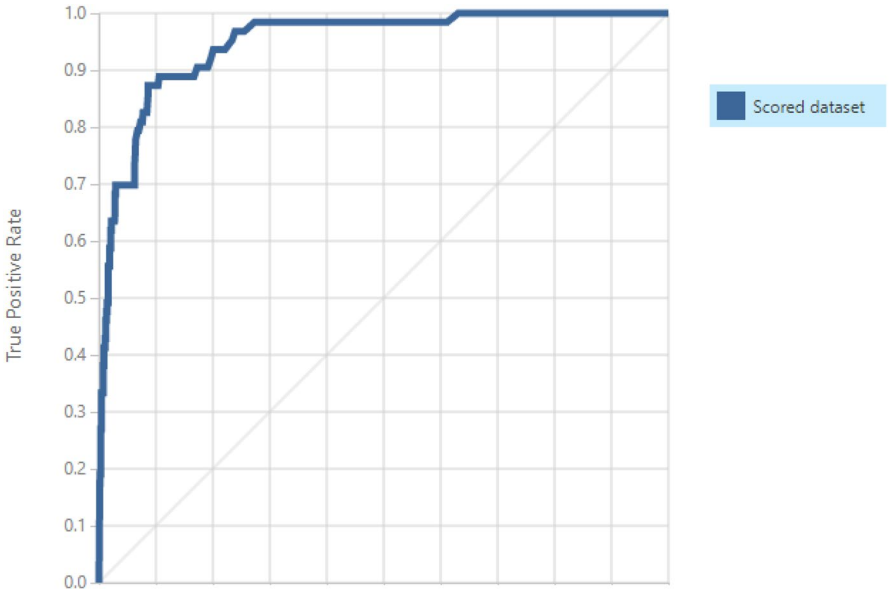


Fig. 2 Legitimacy ROC Curve

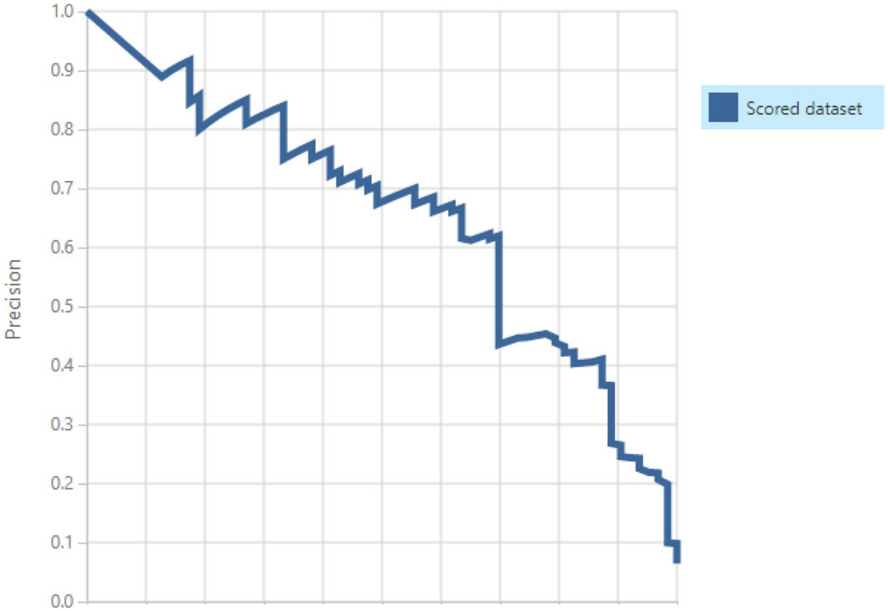


Fig. 3 Legitimacy precision/Recall curve

the number of true positives divided by the sum of the true positives and false positives [4]. It describes how good a model is at predicting the positive class. Precision is referred to as the positive predictive value. Recall is calculated as the ratio of the number of true positives divided by the sum of the true positives and the false negatives. Recall is the same as sensitivity.

C. Lift Curve

The third type of analysis is done using the Lift Curve. The Lift Curve as defined provides the measure to determine the effectiveness of the classifier model so generated [5]. It is the ratio of the result obtained with or without the classifier model. It is the curve on the graph of the population threshold and the rate of positive responses received. The graph consists of a baseline in the middle and the performance of the classifier is evaluated based on the lift curve formed on either the upper side of the baseline or on the lower side.

If it is on the upper side, then it is considered good and the area under the curve is measured, which in this case would be more. If the curve is under the baseline, then it is not a good classifier. The lift curve is seen in Fig. 4.

D. Evaluation Metrics

The final set of statistics used to analyse the machine learning algorithms are five evaluation metrics. For this analysis, the positive label is '*t*,' the negative

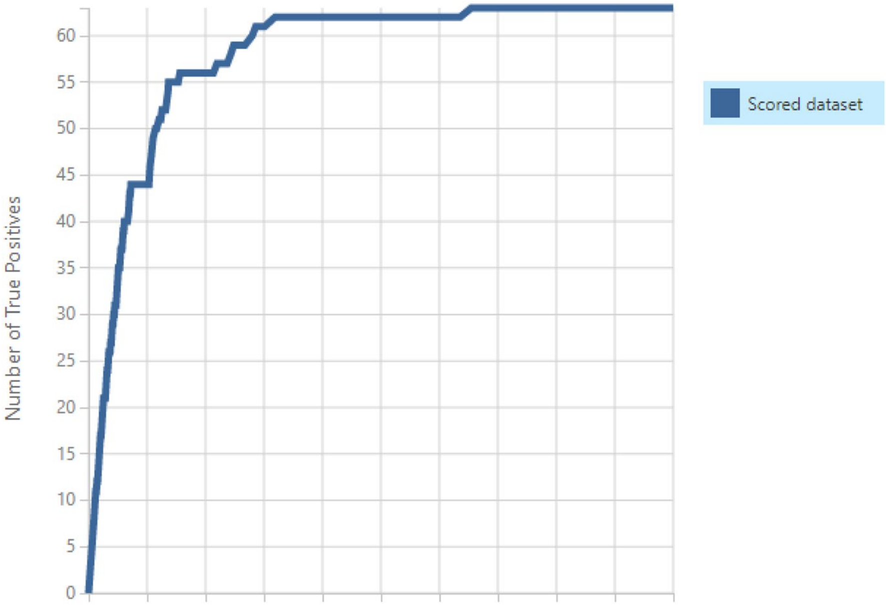


Fig. 4 Legitimacy lift curve

label is ‘*f*,’ and the threshold value is 0.5 for all methods analysed. Accuracy or proportion of correctness is defined by [29] as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{P + N}, \tag{4}$$

where TP identifies the number of true positives generated by the model, TN is the number of true negatives, *P* is the total of positives and *N* is the total of negatives.

Precision or proportion of trueness is defined by [29] as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

where FP stands for the false positives of the model.

Recall or true positive rate is defined by [29] as:

$$\text{Recall} = \frac{\text{TP}}{P} \tag{6}$$

The *F*1 Score or the harmonic mean of precision and recall is defined by [33] as:

$$F1\text{Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{7}$$

The Area Under the Curve (AUC) value as defined by [1] provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has a 1.0 AUC.

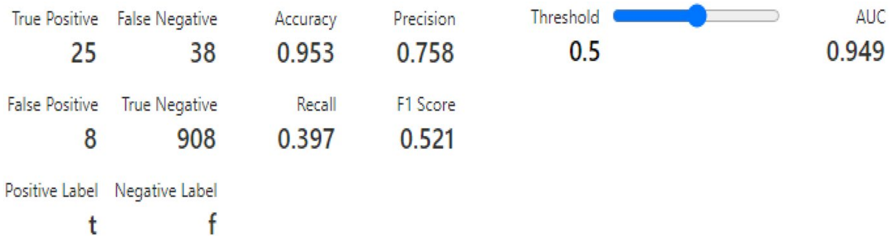
The above metrics are calculated for the Legitimacy ensemble methodology and the results are seen in Fig. 5.

The ROC graph above illustrates when the True Positive Rate (TPR) or Sensitivity is plotted against the True Negative Rate or the Selectivity. According to [35], the closer this curve is to the upper left corner; the better the classifier's performance (that is maximizing the true positive rate while minimizing the false positive rate). From the plot, it can be seen that the True Positive Rate increases rapidly when the False Positive Rate (FPR) is at zero. The ROC curve then continues to plateau near to or equal to a rate of one as the False Positive Rate increases to one. As a result, from this plot, the predictive power of the model chosen in this experiment is seen to be of high performance.

In the Precision/Recall curve, the precision is plotted against the recall. At recall equal to zero, the precision of the experimental procedure begins to decrease to zero and then decreases randomly as there is an increase in the recall. This means that for 100% precision the recall is zero. That further indicates that the model correctly classifies all of the positives at that point in time. As the recall increases to one, the precision shows that it decreases correspondingly. From this analysis, it can be deduced that the ensemble model has a high predictive performance.

From the analysis of the Lift curve, it can be seen that the chosen model increases to 63 true positives quickly as the positive rate increases to one. For the top 65% of predictions, the number of positives fluctuates whilst increasing. After that, the next 35% are equal to sixty-three. This illustrates that for the experimental procedure, there exists the greatest amount of agreement between the actual values and the predicted values. As such, the experimental procedure and the ML model performs well at the task of credibility-fake news detection in a post-disaster scenario.

The results show that the experimental procedure performs well at the task of prediction. The Accuracy of the predictive algorithm was seen to be at 95.3%. The experiment had a 75.8% Precision and a 39.7% Recall. The *F*1-Score was at 52.1% and the AUC value was at 0.949. These results show that the predictive model was highly accurate at the task of credibility-based fake news prediction as well as being highly precise for the same task. This can also be identified by the fact that 25 out of the 63 positive values were rightly classified giving the precision value of 75.8%



True Positive 25
False Negative 38
Accuracy 0.953
Precision 0.758
Threshold 0.5
AUC 0.949
False Positive 8
True Negative 908
Recall 0.397
F1 Score 0.521
Positive Label t
Negative Label f

Fig. 5 Legitimacy evaluation metrics

and only 38 were misclassified giving the recall of 39.7%. For the negative values, only 8 values were misclassified with the majority of 908 being classified correctly leading to such a high accuracy value. The accuracy value reports the state of the trueness of the model and measures the number of true positives and true negatives predicted by the model versus that of the total number of positives and negatives in the test dataset. As such, given the high number of true negatives and the 25 true positives predicted out of the total 63 positives, the accuracy of Legitimacy is seen to be extremely high and good.

6 Comparative Performance

Legitimacy is a model built upon the combination of both a Two-Class Boosted Decision Tree and a Two-Class Neural Network. As such, it is necessary to perform a comparative analysis of the performance of the Legitimacy model against that of the traditional Two-Class Boosted Decision Tree and the Two-Class Neural Network.

6.1 Two-Class Boosted Decision Tree

The performance of the Two-Class Boosted Decision Tree is evaluated using the same four methodologies that are used to evaluate the Legitimacy ensemble, namely the ROC curve, the Precision/Recall curve, the Lift curve, and evaluation metrics. Also, the same dataset is used to evaluate the performance of the Two-Class Boosted Decision Tree to provide an objective analysis.

The performance of the Two-Class Boosted Decision Tree as evaluated by the ROC curve is seen in Fig. 6 below. As can be seen, the Two-Class Boosted Decision Tree performs better in comparison to that of the Legitimacy model. This can be seen from the ROC curve below, which rises immediately and continues almost vertically upwards until it reaches a True Positive Rate of approximately 0.93. This vertical rise illustrates the power of the Two-Class Boosted Decision Tree since it shows that the model is extremely positive for low values of falseness. From that point, the curve plateaus for increasing False Positive Rate values. This illustrates the superior performance of the Two-Class Boosted Decision Tree, because though the False Positive Rate increases the value of the True Positive Rate holds steady and does not change. This means that the model is extremely good at identifying the True positives.

The performance of the Two-Class Boosted Decision Tree as evaluated by the Precision/Recall curve is seen in Fig. 7 below. As can be seen, the Two-Class Boosted Decision Tree performs better in comparison to that of the Legitimacy model because the Precision/Recall curve seen below starts immediately and continues almost vertically downwards for a Recall of approximately 0.00. From that point, the curve continues erratically upwards for increasing Recall values. This illustrates the superior performance of the Two-Class Boosted Decision Tree.

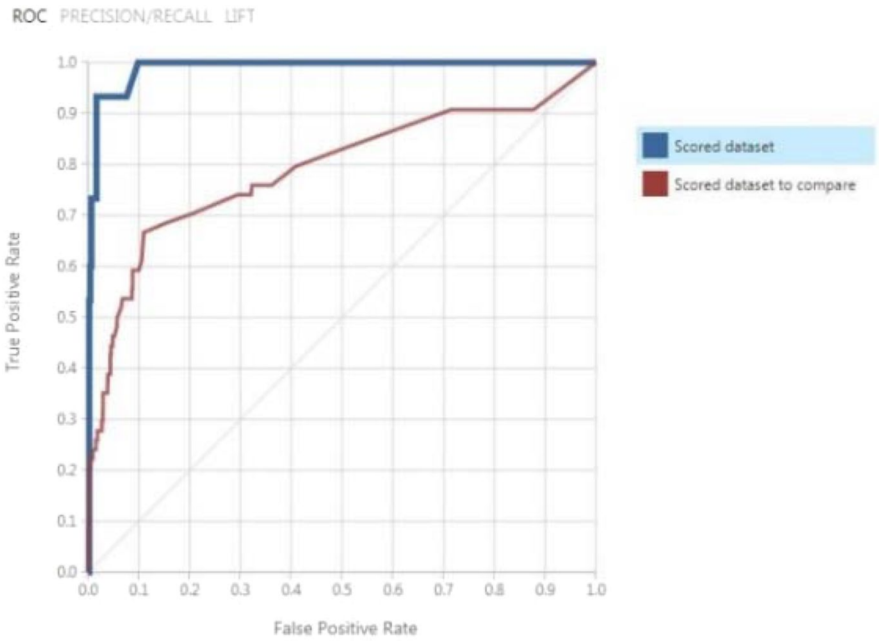


Fig. 6 Two-class boosted decision tree ROC curve

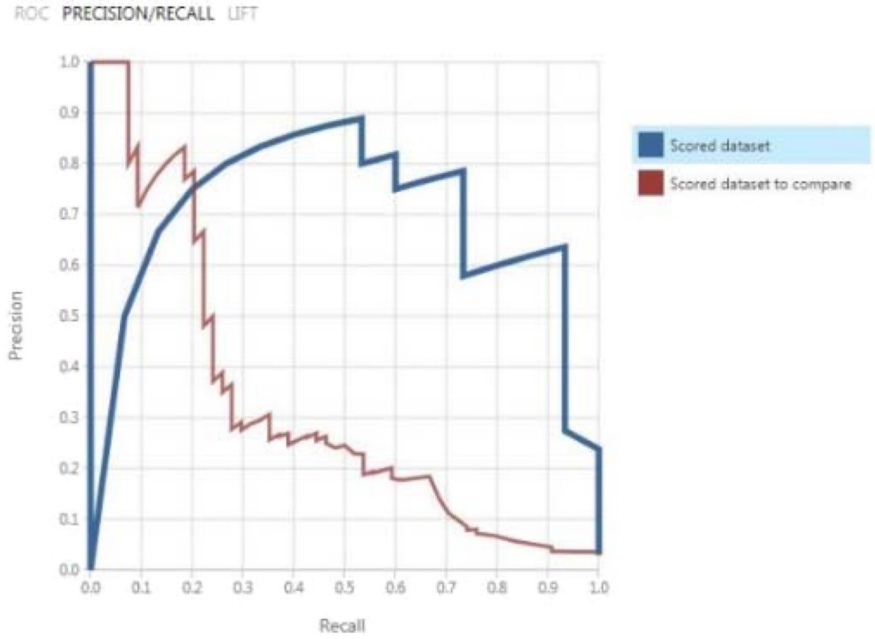


Fig. 7 Two-class boosted decision tree precision/Recall curve

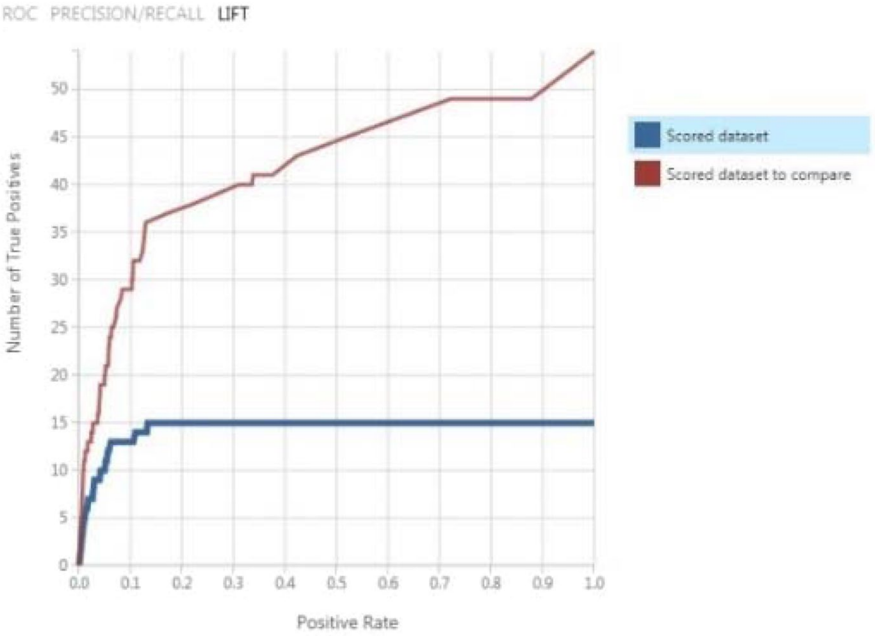
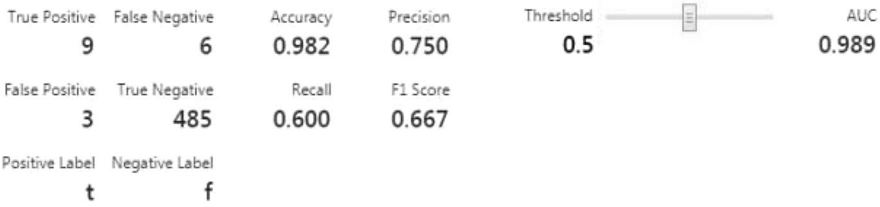


Fig. 8 Two-class boosted decision tree lift curve



True Positive 9	False Negative 6	Accuracy 0.982	Precision 0.750	Threshold 0.5	AUC 0.989
False Positive 3	True Negative 485	Recall 0.600	F1 Score 0.667		
Positive Label t	Negative Label f				

Fig. 9 Two-class boosted decision tree

The performance of the Two-Class Boosted Decision Tree as evaluated by the Lift curve is seen in Fig. 8 below. As can be seen, the Two-Class Boosted Decision Tree performs better in comparison to that of the Legitimacy model because the Lift curve seen below rises immediately and continues almost vertically upwards until it reaches approximately 15 True Positives. From that point, the curve plateaus for an increasing Positive Rate. This illustrates the superior performance of the Two-Class Boosted Decision Tree.

The performance of the Two-Class Boosted Decision Tree as evaluated by the evaluation metrics is seen in Fig. 9 below. As can be seen, the Two-Class Boosted Decision Tree performs better in comparison to that of the Legitimacy model because the Accuracy of the predictive algorithm was at 98.2%. The experiment had 75.0% Precision and 60.0% Recall. The $F1$-Score was at 66.7% and the AUC value was at 0.989. This illustrates the superior performance of the Two-Class Boosted Decision Tree.

6.2 Two-Class Neural Network

The performance of the Two-Class Neural Network is evaluated using the same four methodologies that were used to evaluate the Legitimacy ensemble, namely the ROC curve, the Precision/Recall curve, the Lift curve, and evaluation metrics. Also, the same dataset was used to evaluate the performance of the Two-Class Neural Network to provide an objective analysis.

The performance of the Two-Class Neural Network as evaluated by the ROC curve is seen in Fig. 10 below. As can be seen, the Two-Class Neural Network performs better in comparison to that of the Legitimacy model, because the ROC curve seen below rises immediately and continues almost vertically upwards until it reaches a True Positive Rate of approximately 0.8. This vertical rise illustrates the power of the Two-Class Neural Network since it shows that the model is extremely positive for low values of falseness. From that point, the curve gradually increases for increasing False Positive Rate values. This illustrates the superior performance of the Two-Class Neural Network, because as the False Positive Rate increases, the value of the True Positive Rate also increases, so that the model is extremely good at identifying the True positives.

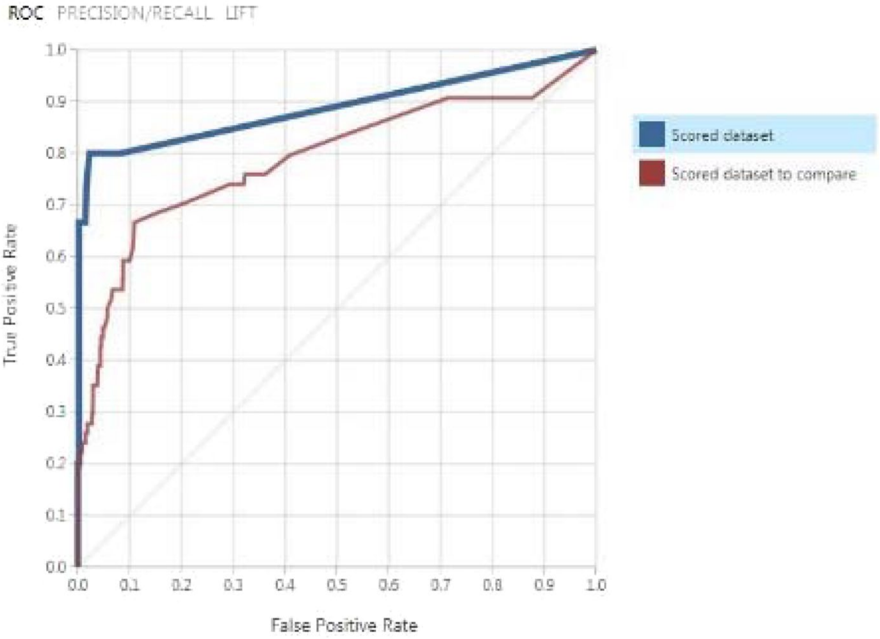


Fig. 10 Two-class neural network ROC curve

The performance of the Two-Class Neural Network as evaluated by the Precision/Recall curve is seen in Fig. 11 below. As can be seen, the Two-Class Neural Network performs better in comparison to that of the Legitimacy model because the Precision/Recall curve seen below, starts immediately horizontal and continues in that direction for a Recall of approximately 0.20. From that point, the curve continues erratically downwards for increasing Recall values. This illustrates the superior performance of the Two-Class Neural Network.

The performance of the Two-Class Neural Network as evaluated by the Lift curve is seen in Fig. 12 below. As can be seen, the Two-Class Neural Network performs better in comparison to that of the Legitimacy model because the Lift curve seen below rises immediately and continues almost vertically upwards until it reaches approximately 13 True Positives. From that point, the curve slightly increases for an increasing Positive Rate. This illustrates the superior performance of the Two-Class Neural Network.

The performance of the Two-Class Neural Network as evaluated by the evaluation metrics is seen in Fig. 13 below. As can be seen, the Two-Class Neural Network performs better in comparison to that of the Legitimacy model because the Accuracy of the predictive algorithm was at 97.6%. The experiment had 100.0% Precision and 20.0% Recall. The $F1$-Score was at 33.3% and the AUC value was at 0.896. This illustrates the superior performance of the Two-Class Neural Network.

Though the results from the individual models are seen to be superior, utilising an ensemble model has many more advantages that make it useful. As explained

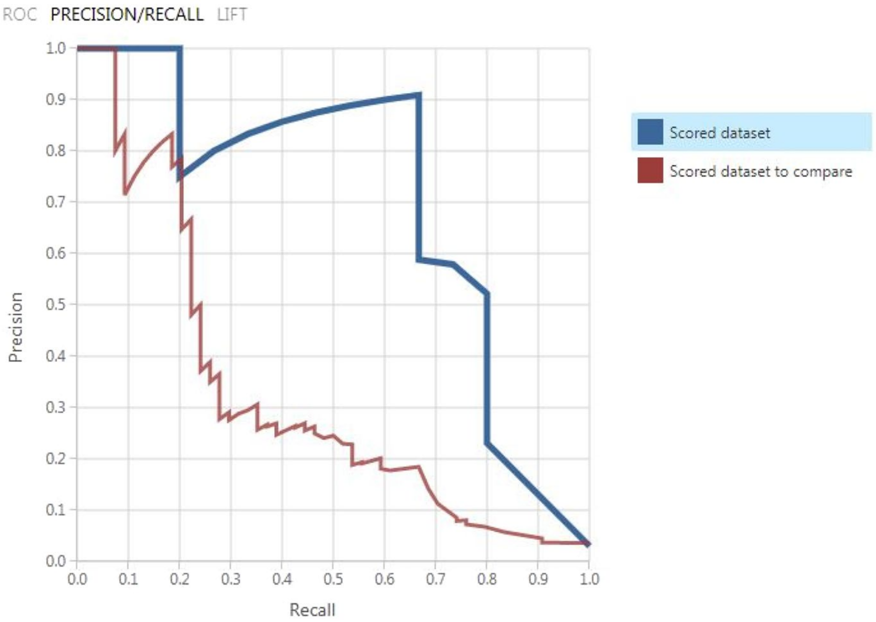


Fig. 11 Two-class neural network precision/Recall curve

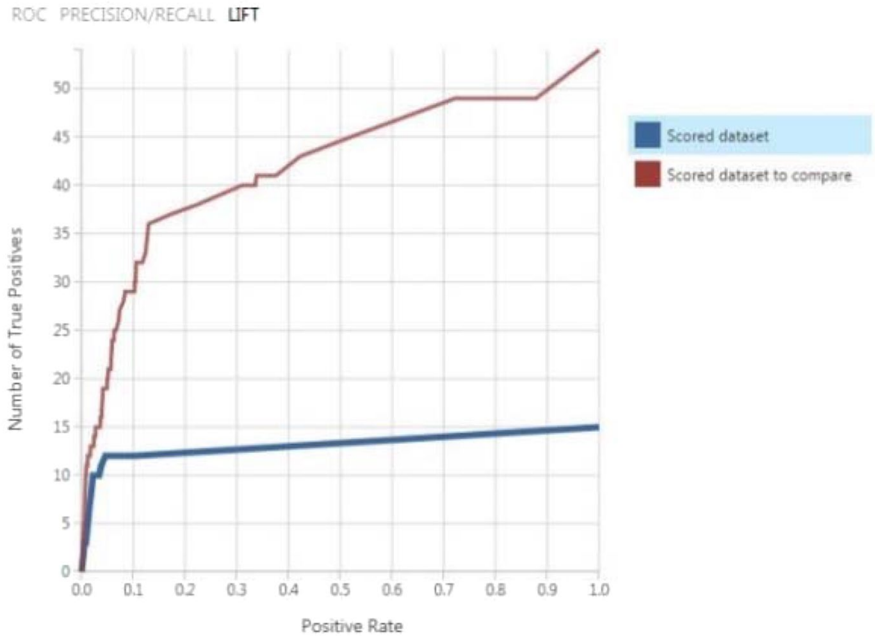
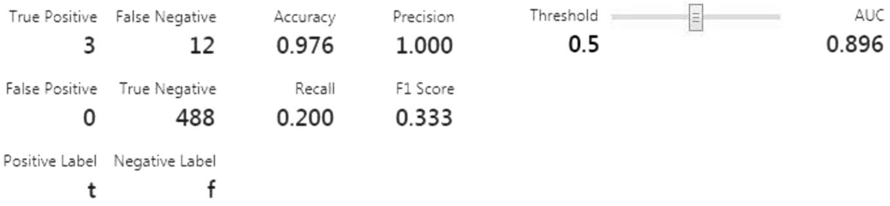


Fig. 12 Two-class neural network lift curve



True Positive 3 | False Negative 12 | Accuracy 0.976 | Precision 1.000 | Threshold 0.5 | AUC 0.896

False Positive 0 | True Negative 488 | Recall 0.200 | F1 Score 0.333

Positive Label t | Negative Label f

Fig. 13 Two-class neural network

in [25], there are several reasons why ensemble methods often improve predictive performance:

- *Overfitting avoidance* When just a small amount of data is available, a learning algorithm is prone to finding many different hypotheses that predict all of the training data perfectly while making poor predictions for unseen instances. Averaging different hypothesis reduces the risk of choosing an incorrect hypothesis and therefore, improves the overall predictive performance.
- *Computational advantage* Single learners that conduct local searches may get stuck in local optima. By combining several learners, ensemble methods decrease the risk of obtaining a local minimum.

- *Representation* The optimal hypothesis may be outside the space of any single model. By combining different models, the search space may be extended and hence, a better fit to the data space is achieved.

Another benefit of the use of the ensemble can be seen from the data features perspectives. The data in current use is composed of 14 input data features and a single class variable. After thorough analysis and experimentation, the dataset can be separated into two distinct subsets namely the demographic and the social behaviour subsets. The social behaviour subset had 8 data features while the demographic subset had 6 features. Using the same type of experimental analysis, it has been discovered that the Two-Class Boosted Decision Tree works best with the social behaviour features while the Two-Class Neural Network works best with the demographic. If either of these models is utilised solely, only a segment of the dataset would be fed into the system. Therefore, the prediction would be based on a subset of the dataset.

With the use of the ensemble model, the Two-Class Boosted Decision Tree and the Two-Class Neural Network are used concurrently to train and predict. As such, using the ensemble, all of the data features are utilised. Hence, the full gamut of features is used to train the model and is used for prediction. This builds for a more robust and versatile prediction versus that of the prediction provided by each of the individual prediction models.

7 Scalability Analysis

The Legitimacy ensemble model has shown resilience as it relates to the size of the dataset increasing with additional data records. To evaluate this result, 17,551, 100,577 and 109,710 synthetic data records were gathered from a simulation environment, respectively. As seen in Table 2, as the size of the dataset grew, the accuracy of the prediction held at the same level. The number of data features was kept constant to ensure that variance in this size did not cause any difference in the predictability of the model.

As can be seen from Table 2, with a dataset of 17,551, the accuracy of the model was 96.9% with 193 false negatives and 5950 true negatives. This is based upon the dataset of 17,551 with a 65%/35% split. With 17,551 data records, the precision is 100% while the Recall and $F1$-Score was 0%, respectively. The AUC value for the 17,551 data records was 0.643.

Table 2 Scalability evaluation metrics

#Nodes	TP	FP	FN	TN	Accuracy	Precision	Recall	$F1$ Score	AUC
17,551	0	0	193	5950	96.9%	100%	0%	0	0.643
100,577	0	0	1227	33,975	96.5%	100%	0%	0	0.5
109,710	0	0	1382	37,016	96.4%	100%	0%	0	0.5

As can be seen from Table 2, with a dataset of 100,577, the accuracy of the model was 96.5% with 1227 false negatives and 33,975 true negatives. This is based upon the dataset of 100,577 with a 65%/35% split. With 100,577 data records, the precision is 100% while the Recall and *F*1-Score was 0%, respectively. The AUC value for the 100,577 data records was 0.500.

As can be seen from Table 2, with a dataset of 109,710, the accuracy of the model was 96.4% with 1382 false negatives and 37,016 true negatives. This is based upon the dataset of 109,710 with a 65%/35% split. With 109,710 data records, the precision is 100% while the Recall and *F*1-Score was 0%, respectively. The AUC value for the 109,710 data records was 0.500. As such, it can be seen that as the number of data records increases the accuracy of the prediction of fake news based on credibility features remained mostly constant as well. This means that the ensemble model is scalable and capable of handling varying numbers of data records.

8 Limitations

Though the proposed Legitimacy Ensemble model has been shown to perform well at the task of credibility-based fake news detection, the following are some limitations to the proposed work.

1. In this study, the dataset is imbalanced, which affects the ML classifiers and skews the results in one direction. This dataset however, is modelled on real life and as such though it is skewed, it reflects reality.

2. In the larger datasets, the mode of each feature is used to fill in missing values for their feature values. The mode is utilised since the data features are not numeric and hence, more appropriate replacement strategies like the mean cannot be used.

3. The dataset used in this study is limited in size, which affects the result of the classifiers, and there is a need to increase the size of the dataset for better results. Though the largest dataset contains 109,710 data records, a larger number closer to 1 million data records is optimal.

4. The Legitimacy Ensemble model is built for the primary purpose of credibility-based fake news detection. As such the model works well with features related to credibility-based fake news detection. It has not been tested on features in other disciplines and as such may require tweaking to function well in such areas.

9 Conclusion

This research used Legitimacy, an ensemble machine learning algorithm for fake news detection and prediction. Legitimacy combined a Two-Class Boosted Decision Tree and a Two-Class Neural Network using a pseudo-mixture-of-experts methodology. The combination of models was accomplished by the use of Logistic Regression. This investigation also performed an experimental analysis of the algorithm using a fixed dataset. These experiments were conducted using the AzureML environment. The experiments were analysed using the ROC, Precision/Recall and Lift curves along with the Accuracy, Precision, Recall, *F*1 Score and AUC values. It was noted that the proposed Legitimacy model performed well with an accuracy of 96.9%. From the experiments performed and the results obtained the Legitimacy ensemble learning model performed excellently. The performance of Legitimacy was compared with that of the base models, namely a Two-Class Boosted Decision Tree and a Two-Class Neural Network. From this analysis, it was found that though the ensemble performed slightly worse than that of each individual model, it was more robust and versatile than each of the ML models. Hence, it can be concluded that based on our selected datasets, the Legitimacy ensemble learning model is an appropriate method suited for detecting and predicting Credibility-Based Fake News. Future work in this area involves the utilisation of further datasets to evaluate the resilience of the Legitimacy model. Also, further research will involve the development of feature selection algorithms that optimize the number of features used in Credibility-Based Fake News Detection.

Funding The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. "Classification: ROC Curve and AUC | Machine Learning Crash Course." Google. Google. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. Accessed 6 June 2020.
2. Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938.
3. Ahmad, I., Yousaf, M., Yousaf, S., & Ovais Ahmad, M. (2020). Fake news detection using machine learning ensemble methods. Complexity. 2020.
4. Brownlee, J. (2019). How to use ROC curves and precision-recall curves for classification in python. Machine Learning Mastery. December 18, 2019. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/. Accessed 1 Jan 2022.
5. Choudhary, R., & Gianey, H. K. (2017). Comprehensive review on supervised machine learning algorithms. In: 2017 International Conference on Machine Learning and Data Science (MLDS), pp. 37–43. IEEE, 2017.
6. Collins, B., Hoang, D.T., Nguyen, N. T., & Hwang, D. (2020). Fake news types and detection models on social media a state-of-the-art survey. In: Asian Conference on Intelligent Information and Database Systems, pp. 562–573. Springer, Singapore, 2020.
7. Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 1–14.
8. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760–772.
9. Dev, V. A., & Eden, M. R. (2019). Formation lithology classification using scalable gradient boosted decision trees. *Computers & Chemical Engineering, 128*, 392–404.
10. Elhadad, M. K., Li, K. F., & Gebali, F. (2019). Fake news detection on social media: a systematic survey. In: 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), pp. 1–8. IEEE, 2019.

11. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861–874.
12. Gaonkar, S., Itagi, S., Chalippatt, R., Gaonkar, A., Aswale, S., Shetgaonkar, P. (2019). Detection of online fake news: A Survey. In: 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1–6. IEEE, 2019.
13. Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Zada Khan, W. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems, 117*, 47–58.
14. Hakak, S., Khan, W. A., Bhattacharya, S., Thippa Reddy, G., & Raymond Choo, K.-K. (2020). Propagation of fake news on social media: challenges and opportunities. In: International Conference on Computational Data and Social Networks, pp. 345–353. Springer, Cham, 2020.
15. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.
16. Martens, J. (2020). "Machine Learning Studio (Classic) Documentation - Azure." Machine Learning Studio (classic) documentation - Azure | Microsoft Docs. Accessed April 22, 2020. https://docs.microsoft.com/en-us/azure/machine-learning/studio/. Accessed 22 Apr 2020.
17. Kaliyar, R. K., Goswami, A., Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. In: 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 103–107. IEEE, 2019
18. Kazllarof, V., Karlos, S., & Kotsiantis, S. (2019). Active learning rotation forest for multiclass classification. *Computational Intelligence, 35*(4), 891–918.
19. Khan, H., Asghar, M. U., Asghar, M. Z., Srivastava, G., Reddy Maddikunta, P. K., Gadekallu, T. R. (2021). Fake review classification using supervised machine learning. In: International Conference on Pattern Recognition, pp. 269–288. Springer, Cham, 2021.
20. Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets. *SMU Data Science Review, 1*(3), 9.
21. Ramkissoon, A.N., & Goodridge, W. (2021). Legitimacy: An ensemble learning model for credibility based fake news detection. In: 2021 International Conference on Data Mining Workshops (ICDMW), pp. 254–261. IEEE, 2021.
22. Ramkissoon, A. N., Mohammed, S., & Goodridge, W. (2021). Determining an optimal data classification model for credibility-based fake news detection. *The Review of Socionetwork Strategies, 15*(2), 347–380.
23. Richard, & Lovell, J. (2020). The war on fake news: College of Communication. College of Communication The War on Fake News Comments. Accessed 5 Feb 2020.
24. Roy, A., Basak, K., Ekbal, A., Bhattacharyya, P. (2018). A deep ensemble framework for fake news detection and classification. arXiv preprint arXiv:1811.04670.
25. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1249.
26. 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. https://doi.org/10.1145/3137597.3137600
27. Shu, K., Wang, S., Liu, H. (2018). Understanding user profiles on social media for fake news detection. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 430–435. IEEE, 2018.
28. Sugiyama, M. (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.
29. Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodoloski zvezki, 3*(1), 89.
30. Younus Khan, J., Khondaker, T. I., Iqbal, A., & Afroz, S. (2019). A Benchmark study on machine learning methods for fake news detection. arXiv preprint arXiv:1905.04749.
31. Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems, 23*(8), 1177–1193.
32. Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management, 57*(1), 102107.
33. Zhang, D., Wang, J., & Zhao, X. (2015). Estimating the uncertainty of average F1 scores. In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 317–320.
34. Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management, 57*(2), 102025.
35. Zhou, X., & Zafarani. (2018). Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.