A Fake Profile Detection Model Using Multistage Stacked Ensemble Classification

Swetha Chikkasabbenahalli Venkatesh[1,*], Sibi Shaji[1], Balasubramanian Meenakshi Sundaram[2]

[1]School of Computational Sciences & IT, Garden City University, Bangalore, India

[2]Department of Computer Science & Engineering, New Horizon College of Engineering, Bangalore, India

Received 15 December 2023; received in revised form 11 January 2024; accepted 12 January 2024

DOI: https://doi.org/10.46604/peti.2024.13200

Abstract

Fake profile identification on social media platforms is essential for preserving a reliable online community. Previous studies have primarily used conventional classifiers for fake account identification on social networking sites, neglecting feature selection and class balancing to enhance performance. This study introduces a novel multistage stacked ensemble classification model to enhance fake profile detection accuracy, especially in imbalanced datasets. The model comprises three phases: feature selection, base learning, and meta-learning for classification. The novelty of the work lies in utilizing chi-squared feature-class association-based feature selection, combining stacked ensemble and cost-sensitive learning. The research findings indicate that the proposed model significantly enhances fake profile detection efficiency. Employing cost-sensitive learning enhances accuracy on the Facebook, Instagram, and Twitter spam datasets with 95%, 98.20%, and 81% precision, outperforming conventional and advanced classifiers. It is demonstrated that the proposed model has the potential to enhance the security and reliability of online social networks, compared with existing models.

Keywords: fake profile, online social networks, stacked ensemble, imbalanced dataset, cost-sensitive learning

1. Introduction

The proliferation of online social networks (OSN) such as Instagram, Facebook, Twitter, and YouTube has completely altered how people connect, communicate, and engage with one another in the digital era [1]. These platforms, which have rooted themselves firmly in the contemporary social structure, enable the interchange of information on a scale never before feasible. These platforms make it substantially easier to exchange ideas and cultivate ties between individuals. However, along with the rapid growth of OSNs, there has been a simultaneous rise in fraudulent behavior, most notably in the form of fake profiles. Fake profiles, which are frequently created to cause harm, pose a threat to the security, reliability, and integrity of online communities [2]. It is a significant challenge to identify fake profiles that are present in OSNs and to take measures to reduce their impact.

By carefully copying the characteristics of legitimate users, these fake profiles are designed to be difficult to identify using common methods [3]. The severity of the problem is highlighted by the fact that these platforms can enable a wide variety of unethical behaviors, such as the dissemination of false information, cyberbullying, phishing attacks, and other types of illegal behavior [4]. As a result, developing solutions to handle this diverse challenge requires the adoption of new, resilient ways that are competent at distinguishing between genuine and fake profiles in the digital domain.

* Corresponding author. E-mail address: cvswetha1987@gmail.com

English language proofreader: Yen-Chun Hsieh

Over the last several years, numerous strategies and procedures have been developed for fake profile detection [5]. These approaches encompass a wide variety of methodologies, including machine learning (ML) [6], data mining (DM), natural language processing (NLP) [7], and behavioral analysis [8], among others. Some of the common ML classifiers are used in the existing works, such as decision trees (DTs) [9], K-nearest neighbors (KNN) [10], logistic regression (LR) [11], random forest (RF) [8], naive Bayes (NB) [12], extreme gradient boosting (XGBoost) [13], and support vector machines (SVMs) [14] for classifying profile data in OSNs. While existing techniques for detecting fake profiles have contributed significantly to the literature, further work is needed to design robust and diverse models capable of successfully addressing the difficulties, especially those introduced by unbalanced datasets [15]. The imbalance in the number of real profiles relative to fake profiles might reduce the reliability of detection algorithms. High accuracy has been difficult to achieve using traditional methods, especially when dealing with the difficulties of real social networks.

Furthermore, researchers are exploring various feature sets to improve fake profile identification. Thus, focusing on the most important features is crucial for efficiency and accuracy, and a lack of feature selection can lead to reduced model performance, higher computing costs, and decreased interpretability. Thus, accurately identifying fake profiles within the context of social media platforms is a contemporary challenge. These profiles typically have characteristics and behaviors that are similar to those of real individuals, but they are motivated by malicious objectives. To keep online communities safe and secure, these profiles must be identified instantly. However, traditional detection methods struggle with dataset imbalance, resulting in suboptimal performance due to genuine profiles outnumbering fake ones, highlighting the need for novel fake profile detection algorithms.

Thus, the objectives of this research are twofold: First, it aims to propose a novel strategy to improve the effectiveness of identifying fake profiles in OSNs by overcoming the difficulties posed by the uneven distribution of classes in the datasets. Specifically, it achieves this objective through the development of a novel multi-stage stacked ensemble classification model for detecting fake profiles. This model incorporates innovative concepts like novel feature selection, ensemble learning, and cost-sensitive learning. Second, the research seeks to evaluate the proposed multi-stage stacked ensemble classification model to validate its efficiency. This is achieved through extensive analysis, which shows the effectiveness of the proposed model in comparison with other traditional and state-of-the-art classifiers. The objectives are mapped to the working hypothesis, which is fundamental for guiding the research. The formulated hypothesis is that the proposed multi-stage stacked ensemble classification model will significantly improve the accuracy of fake profile detection. The contributions of the research work are summarized below:

(1) A Chi-squared feature-class association-based feature selection and SMOTE for data pre-processing

(2) A stacked ensemble classification with cost-sensitive learning

(3) A meta-cost-based neural networks

The paper is organized as follows: Section 2 discusses the related works. In Section 3, the proposed multistage stacked ensemble classification model is explained. It includes a chi-squared feature-class association model for feature selection, stacked ensemble classification with the base learner phase, and a meta-classifier phase. Section 4 discusses the experimental and result analysis, including the dataset used, performance indicators, and result analysis. Finally, the conclusion section concludes the paper.

2. Related Works

Several studies have been conducted to review techniques for detecting compromised accounts, providing a foundation for future research in this field and highlighting the need for comprehensive analysis [3]. The survey conducted by Drury et al. [4] aims to document social media crimes, taxonomize similar crimes, and provide suggestions for further research using

publicly available datasets. A literature review explored recent advancements in ML techniques for bot detection and classification on social media platforms like Facebook, Instagram, LinkedIn, Twitter, and Weibo [15]. Like any other application, feature selection in classification problems is significant in high-dimensional data contexts. Bahassine et al. [7] proposed a novel Arabic text classification method using ImpCHI and SVM classifiers, enhancing performance by 90.50% compared to traditional feature selection metrics. A study was carried out using dynamic feature selection and ML algorithms to identify spam users on Twitter due to the rise in malicious activities [10]. Purba et al. [8] used 17 features to identify fake Instagram followers and proposed supervised ML models to classify authentic and fake users, with the RF algorithm providing the highest accuracy.

Research on fake profile detection using ML techniques has gained more attention in recent days. A model was proposed to classify a cluster of accounts as malicious or legitimate based on user-generated text, including patterns within the cluster and comparisons across the user base. Hassan et al. [6] proposed a supervised learning algorithm with SVM and RF classifiers to identify fake social media accounts. The results indicated that the model outperforms other techniques for safeguarding networks from online threats. Liang et al. [11] explored spam detection on Sina Weibo, a Chinese microbiological website, using ML methods, revealing that the LR model outperforms NB and J48 DT in terms of precision, recall, F-measure, and area under curve (AUC).

ML algorithms, a cost-sensitive genetic algorithm for automated accounts, and the SMOTE algorithm for unevenness in the fake profile dataset were used in a study to identify Instagram fake and automated accounts [14]. Elyusufi et al. [9] employed DT and NB algorithms to detect fake profiles on social media and classified user profiles into genuine and fake. Sallah et al. [16] proposed an ML architecture for detecting fake Instagram accounts using techniques like bagging and boosting, synthetic minority over-sampling technique (SMOTE), and SHapley Additive exPlanations (SHAP) values, with a combined accuracy of 96% using XGBoost and RF models. However, most of these studies lack accuracy and extensive evaluation and comparison with state-of-the-art classifiers.

Aydin et al. [17] employed ML techniques to identify fake accounts on social networks like Twitter, with LR being the most effective classification method. A study on Facebook data used DM techniques to detect fake profiles, with ID3 showing the highest accuracy, addressing the issue of misused accounts for malicious activities [18]. Sahoo and Gupta [19] presented an ML method for detecting fake profiles from Facebook and Twitter, utilizing multimedia big data to analyze content and profile features, demonstrating its potential. The prevalence of cybercrime, especially against women, is examined in a study through the use of ML techniques applied to Instagram accounts [20]. The result revealed that RF outperforms LR in detecting fake profiles. Kaushik [21] proposed a new algorithm to detect automated spam accounts and fake profiles on Instagram, achieving precision and accuracy of 93% and 91%, highlighting the need for improved security measures. The author also reviewed the use of various detection methodologies, such as deceptive, predictive, linguistic, and clustering; however, these methods are suitable only for fake news detection and not fake profile detection.

In the field of fake profile identification on OSNs, deep learning has emerged as a potential method. Traditional systems struggle with fake profiles owing to synthesized features, whereas deep learning can learn advanced patterns from raw data. RunFake, a dynamic convolutional neural network (CNN) for malicious account classification, was proposed using a general activation function called RunMax to improve training and testing accuracy [2], and the method produced better results with features involving user profile data. A heterogeneous stacking-based ensemble learning framework was suggested for improving spam detection in social networks. It uses six base classifiers and cost-sensitive learning, enhancing detection rates on imbalanced datasets which serves as a base for the proposed work [5]. A study suggested digital face-processing authentication as a double-factor authentication method for OSN, with deep learning classification achieving 95% accuracy and SVM achieving 97.8% for fake profile detection [12]. The summary of significant existing studies for fake profile detection, along with the obtained results and their limitations, is presented in Table 1.

Table 1 Summary of significant existing studies on fake profile detection

Study	Dataset used	Preprocessing	Results	Limitations
Various classifiers were assessed including RF, neural network, LR, NB, and DT [8]	Real-time Instagram dataset (32,460 samples with 2 classes; 10,441 samples for 4 classes)	The correlation between the features was evaluated	The RF provided improved performance with 91.76%	Lack of accuracy
Classifiers such as DT, RF, and NB were assessed [9]	Real-time Facebook dataset (2816 user accounts)	Random sampling and 4 out of 33 features are selected	The DT offered improved performance with 99.30% accuracy	Lacks accuracy and details of feature selection techniques used
Classifiers like NB, DT, and LR were assessed [11]	Realtime Sina Weibo dataset (8149 users)	-	Improved prediction accuracy of 90.60%	Lacks accuracy and detailed evaluation
AlexNet neural network - Stochastic gradient descent-based deep learning was suggested [12]	Real-time dataset (1207 faces)	Random sampling	Improved detection accuracy with 94.70%	Other profile features need to be considered
LR, NB, SVM, KNN, and neural network classifiers were assessed [12]	Facebook spam dataset (600 profiles) and Instagram fake spammer genuine accounts (696 profiles)	-	SVM and NB offered an increased accuracy of 97.80%	Lacks performance and evaluation with large datasets
Genetic feature selection is proposed with NB, LR, SVM, and neural network classifiers [14]	Real-time Instafake-dataset (1203 user accounts)	Applies SMOTE-NC algorithm for data balancing	Improved accuracy of 94% with SVM and neural networks	Evaluated with a smaller number of samples
Multimedia big data with multiple classifiers was suggested [19]	Synthesized Facebook dataset (2000 profiles)	Content and profile-based features	Offered increased detection accuracy of 99.52%	Lack of accuracy and comparison with deep learning techniques
LR and RF classifiers were employed [20]	Facebook spam dataset (600 profiles) and Instagram fake spammer (696 profiles)	-	RF offered a higher detection accuracy of 90.80%	Lacks comparison analysis
Deep learning with artificial neural networks (ANN) was suggested [21]	Instagram fake spammer genuine account (576 user accounts)	-	Increased detection accuracy of 91% compared with other classifiers	Does not use complex structural features and fails for short profiles

Most of the prior research has been focused on identifying spam accounts on social networking sites using conventional classifiers and only a few of them used deep learning techniques for effective results. Additionally, there is no significance for feature selection in the synthesized imbalanced datasets that the researchers used for evaluation. Further, in datasets where real profiles outweigh fake ones, standard detection approaches frequently encounter difficulties. The literature review reveals that there is no precise or flawless method for detecting fake accounts. As a result, this serves as a motivation to propose a fake profile detection strategy that extracts significant features, performs data balance, and utilizes meta-learning to deliver enhanced performance.

3. Proposed Method

This section presents the multistage stacked ensemble classification model for fake profile detection in OSNs inspired by Zhao et al. [5]. Generally, ensemble learning is a robust technique in ML that is frequently employed to enhance prediction performance in a wide range of contexts. More specifically, stacked models, also known as stacked ensembles, are a popular ensemble learning technique where multiple base models are trained on a dataset and a meta-model is trained to combine predictions. The meta-model calculates ensemble predictions by weighing predictions from each base model, often

outperforming individual models by leveraging their strengths. Thus, to improve the detection of fake profiles, a novel architecture with a three-level framework is proposed, comprising feature selection, base learner, and meta-learner phases. Fig. 1 depicts the proposed framework for fake profile detection and demonstrates the working procedure of multistage stacked ensemble classification for an imbalanced dataset. Here, the dataset undergoes SMOTE [22], an ML technique employed to address class imbalance in datasets. This helps prevent biased model training and ensures better predictive performance.

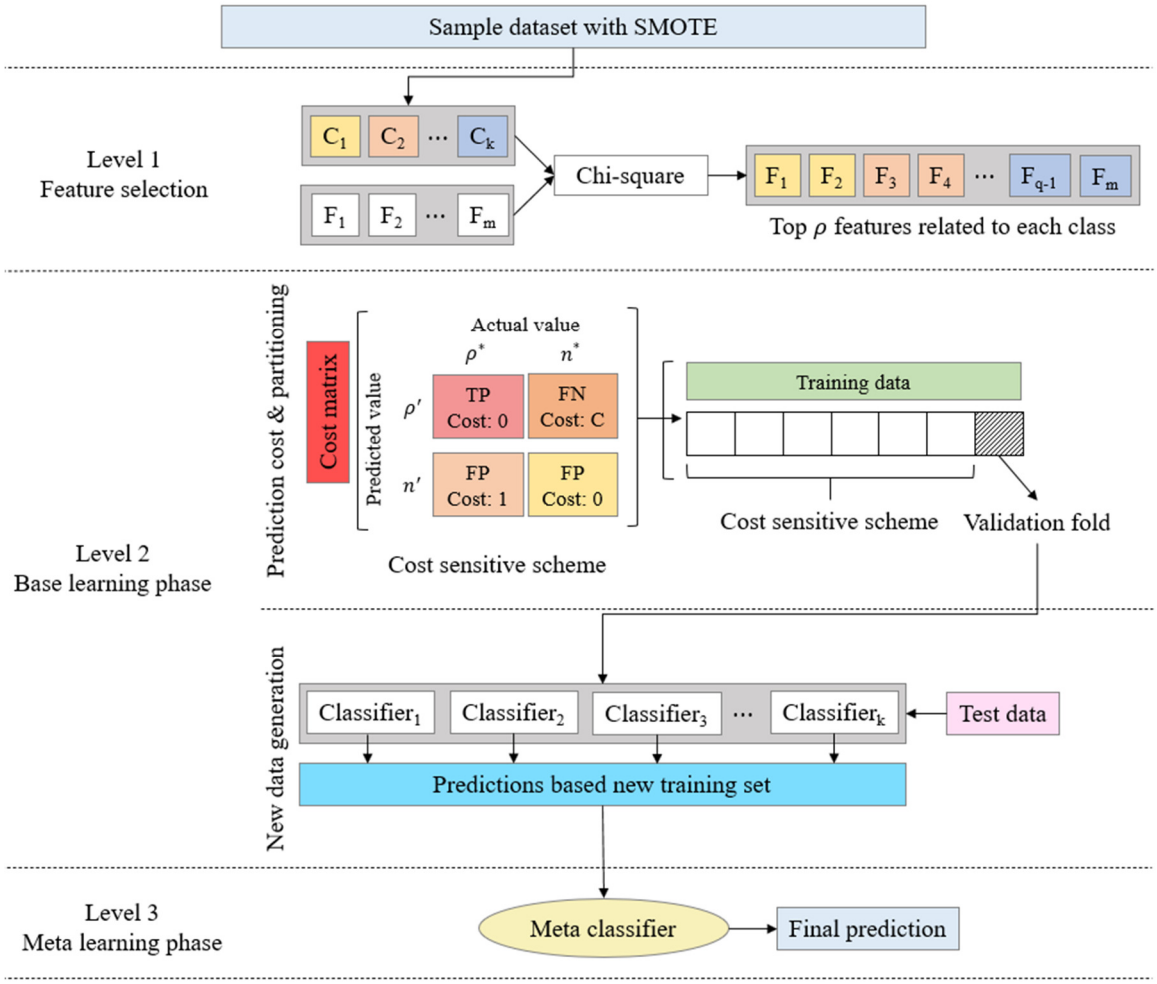

Fig. 1 Proposed fake profile detection framework

3.1. Chi-squared feature-class association model for feature selection

Features selection enhances the prediction rate in classification problems by reducing high-dimensional data since redundant or irrelevant features can cause overfitting and poor performance [7]. Feature selection is crucial for improving the accuracy of the ML models and efficiency in identifying fake profiles while reducing noise and computational costs. Feature selection improves detection efficiency by focusing on key characteristics like posting frequency, profile metadata, user engagement, and content quality, enabling the identification of genuine and fake profiles. Researchers have introduced features to enhance fake profile detection accuracy, but only a small percentage have implemented significant feature selection methods, including correlation analysis [8], Bayesian information criteria [10], genetic algorithm [14], and other filter methods (information gain, chi-square, variance threshold, and more).

The model used in this work independently computes the chi-squared test for each attribute and class variable [23]. In general, the chi-square test (x^2) compares the variances of two distributions and is used in feature selection to determine the significance of a relationship between a feature and a class variable. The chi-squared statistic is calculated for each feature in the dataset and sorted in descending order, with a high value indicating its significance in class label determination. It uses two

values, the observed frequency which is the actual count or number of observations in a given group, and the expected frequency which is the expected number of observations assuming no association between two categorical variables, calculated based on independence. Thus, to calculate the chi-squared for each of a feature's attributes *i*, the final chi-square score for each class *k* is found by aggregating the chi-squared values of all of the feature's attributes. The work selects top-scoring features from each class based on class proportion, to limit the number of features to be selected from each class.

3.2. Stacked ensemble classification for imbalanced dataset

Stacking is a robust ensemble learning technique that effectively detects fake profiles by combining the insights of multiple models. As the model includes a wide array of classifiers, it helps to improve overall accuracy and robustness. Predictive performance is improved by base models, most widely of different types, that capture distinctive aspects of the data [5]. The meta-learner generates final predictions after training on the outputs of these underlying base models. Thus, the proposed model has two stages in this phase of classification. The first stage consists of an ensemble of base classifiers and is called the base learning phase, and the next stage accumulates the results from the base layer for predicting the class and is called the meta-learning phase.

3.3. Base learning phase: cost-sensitive base classifiers

The significance of stacking lies in its capacity to substantially enhance accuracy. Training a meta-learner on the predictions of these diverse base models generates final predictions that are often more precise and resilient than those of individual models. This improved accuracy renders the more effective identification of fake profiles, ultimately reducing the risks associated with false positives and negatives. Furthermore, the adaptability of stacking ensures that detection methods remain effective as OSNs and fake profile strategies evolve. Ultimately, its role extends beyond mere detection; it contributes to fostering trust and safety in OSNs by safeguarding users from the harmful consequences of fake profiles and preserving the credibility of online social communities.

The goal of the base learning phase is to train the base classifiers using the training set, and the metaclassifier is trained using the metadata produced by the base classifiers. As the significance of stacking comes from its ability to improve performance, the selection of suitable base models, efficient use of computational resources, and effective implementation of ensemble methods are crucial for optimal performance. It is often held that a vital component of successful integration is the ability to measure the variety of the underlying individual classifiers [5]. It was reported that diversity and complementarity among the base classifiers are essential for obtaining advanced information for classification.

Thus, in the proposed work on detecting fake profiles in OSN, seven base classifiers that are frequently used in the existing research on fake profile detection and found to be effective are chosen. These classifiers include DT [9], AdaBoost [16], KNN [10], LR [11], RF [8], NB [12], XGBoost [13], and SVM [14]. Each of these classifiers has its strengths in resolving classification problems, especially in fake profile detection. The prediction results from these base classifiers are fed into the meta-learning phase in this stacked ensemble learning approach. However, a cost-sensitive learning strategy is used with these base classifiers to account for the uneven class distribution in the dataset.

The primary use of cost-sensitive learning is at the base classifier level, where each base classifier is trained using a cost-sensitive method that takes the costs of misclassification associated with various classes into account. The user determines the costs associated with solving the classification. It is not possible to train the combining module with prediction results from the underlying classifiers using the entire dataset. Overfitting is a major concern when validating a model, and this is especially true when the same data is used for both training and testing the model. In the proposed staking ensemble, k-fold cross-validation is employed.

3.4. Cost-sensitive learning

In fields like medical diagnosis, fraud detection, and risk assessment, where the consequences of various types of errors vary greatly, cost-sensitive learning is crucial. In an imbalanced classification problem, an ML paradigm known as "cost-sensitive learning" takes into account the variable costs related to various types of misclassifications.

The objective of traditional binary classification is to reduce the classification error (the number of misclassified instances). The cost of such misclassifications, however, might vary greatly depending on the situation in many real-world scenarios. For instance, incorrectly classifying a cancer patient as healthy (false negative, FN) can have considerably more serious implications than incorrectly diagnosing a healthy individual as a cancer patient (false positive, FP). This is addressed by cost-sensitive learning, which incorporates these costs into the learning phase.

Further, different cost matrices are employed as the penalty for incorrect classification in a cost-sensitive classifier to address the issue of unbalanced data learning. The structure of the cost matrix, which shows the relative importance of various types of classification errors and quantifies the costs associated with misclassification, is specified in Table 2, where columns indicate the predicted class and rows indicate the actual class [24].

Table 2 Cost matrix

Actual vs. predicted	Predicted positive	Predicted negative
Actual positive	$C(0,0) = c_{00}$	$C(0,1) = c_{01}$
Actual negative	$C(1,0) = c_{10}$	$C(1,1) = c_{11}$

Conceptually, instances that are correctly classified should not have any penalty, and thus the values will always have zero costs. In the case of fake profile detection, fake profiles are regarded as positive samples, and genuine profiles as negative samples. Therefore, the misclassification cost of c_{01} belonging to the minority class should be higher than that belonging to the majority c_{10} [5, 25]. In cost-sensitive learning, the primary idea is to train the classifier to have a minimum classification error. That is, it classifies the training samples into the class that has the minimum expected cost. The optimal prediction for a given observation, *v* given the values of the predictors, x_v is the class *i* among all *I* classes with the lowest value.

$$\sum_{j=1}^{n} P(j|x_v)C(i,j) \tag{1}$$

In Eq. (1), $P(j|x_v)$ is the conditional probability of class *j* evaluated by the classifier for an observation *v* with predictor vector x_v. $C(i,j)$ is the cost of classifying class *i* when the actual class for observation *v* is *j*. In other words, the cost matrix indicates that the model's training should prioritize lowering the sorts of mistakes that result in greater costs.

3.5. Partitioning

The k-fold cross-validation method is a critical tool for model assessment and ensemble learning in the context of stacking. When performing stacking, numerous basic classifiers are combined into one powerful meta-model, and the performance of this method is improved by the use of k-fold cross-validation [5].

The first step is to fold the training dataset into k equal-sized subgroups. To preserve the ratio of classes, each fold should contain data that has been randomly selected and stratified. The basic classifiers are trained using k-1 of the subsets for each fold. In practice, this implies that the base classifiers are trained k times, each time with a slightly different training subset. After k folds have been used to train base classifiers, the remaining fold is used as a validation set to assess the accuracy of the trained classifiers. This estimates the likelihood of each base classifier's ability to generalize to new data. A total of k sets of predictions from each base classifier across the complete training dataset will be obtained when k iterations are finished (since

each fold serves as the validation set once). These final output features from the base classifiers are fed into a meta classifier. To create the final predictions, the original target labels are used to train a meta-model. Predictions are then made using the stacked model, which is a combination of the predictions from the base classifiers.

Thus, k-fold cross-validation allows base classifiers to learn from different training data subsets, reducing overfitting and improving ensemble generalization to new data. The training of the meta-model on base classifier predictions allows the stacking ensemble to leverage the unique strengths of individual models, enhancing predictive accuracy and robustness.

3.6. Meta-learning phase: meta-cost multilayer perceptron (MLP) classifier

For the meta-learning phase, deep learning with ANN is used with cost-sensitive learning. In general, many neurons, structured in several layers, make up an ANN. Neurons are processing units that receive function-specific data. A formula evaluates numerous input layer weights and propagates the findings to successive layers to calculate neuron values. Thus, it simulates physiological neurons that send information based on connectivity weights. The ANN implementation known as MLP has achieved success in a wide variety of industries. The general framework of MLP has an input layer that receives the raw data or features as its input, one or more hidden layers positioned between the input and output layers, and performs complex transformations on the input data. Finally, the output layer produces the final prediction or output of the network.

Weights are assigned to the connections between neurons. These weights are modified during training to reduce the discrepancy between the predictions of the network and the true targets. To simulate complicated interactions, activation functions add non-linearity to the network. The sigmoid function, tanh, and rectified linear unit (ReLU) are popular activation functions employed for improving efficiency. Data is fed forward through the network for training, errors propagated backward, and weights adjusted iteratively to minimize errors until the model converges or reaches a stopping criterion. The model uses labeled training data to improve predictive accuracy by iteratively adjusting weights and biases, aiming to uncover underlying patterns and relationships. The model can make predictions on new, unseen data by applying learned weights and biases to the input features after training.

However, due to the imbalanced class distribution, instead of using the MLP directly, the cost-sensitive version of the algorithm is used. In ML, meta-cost is a method for addressing the problem of class bias [26]. The training of the model is optimized by redistributing the misclassification costs among the various classes. The minority class (the less common group) is penalized more than the majority class (the more common group). By making this modification, models like MLPs are trained to reduce their bias towards the majority class in favor of producing more accurate predictions for the minority class [27]. Fine-tuning these costs improves the ability of the model to handle unbalanced datasets and identify all classes, resulting in more balanced and accurate predictions across real-world classification scenarios. The pseudocode of the meta-cost algorithm adopted by Domingos [26] and Ghatasheh et al. [27] is summarized and illustrated in Algorithm 1. The workflow of the proposed work is depicted in Fig. 2.

Algorithm 1: Meta-cost algorithm

```
Input: Training set S, MLP learning classifier, C – cost matrix, m – no. of resamples to generate, n – no. of examples in
each resample, p – True iff MLP produces class probabilities, q – True iff all resamples are to be used for each example
Procedure MetaCost (S, MLP, C, m, n, p, q)
Begin
  For i = 1 to m do
    Si – subset of S with n examples
    Mi – Model produced by applying MLP to Si
  For each example x in S
    For each class j
      Apply Eq. (1) for optimal prediction
```

If p is True then
$P(j|x, M_i)$ is produced by M_i
Else
$P(j|x, M_i) = 1$ for the class predicted by Mi for x, and 0 for all others
If q is True then
i ranges over all M_i
Else
i ranges over all M_i such that $x \notin S_i$
Let x's class = $\text{argmin}_i \Sigma_j P(j|x)C(i,j)$
Let M = Model produced by applying L to S
Return M
End Procedure

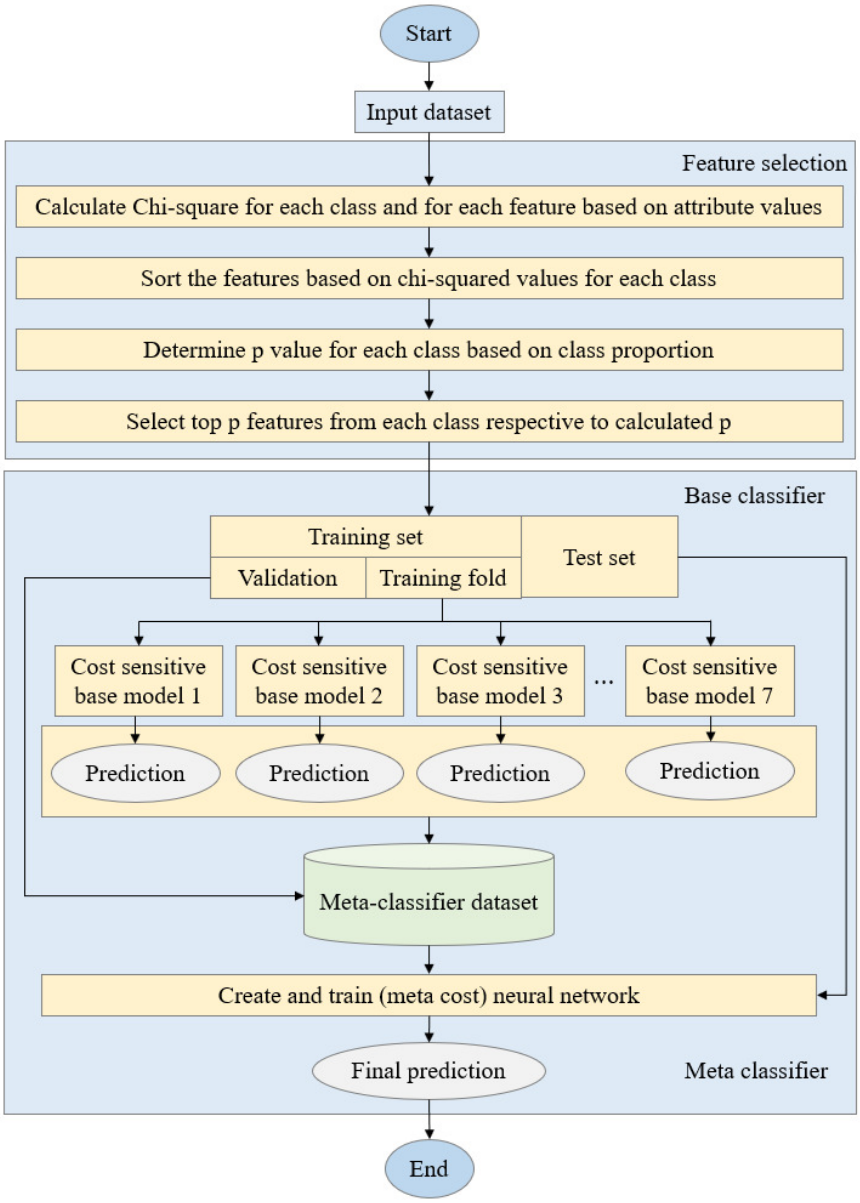

Fig. 2 Workflow diagram for the proposed model

4. Experiment and Analysis

This paper examines the effectiveness of heterogeneous stacking-based ensemble learning methods in fake profile detection on social networks, comparing their performance with other algorithms. This section provides a detailed description of the dataset, metrics used to assess results, experimental design, and discussion of the experimental results.

4.1. Dataset description

The three widely used and publicly accessible Instagram, Facebook, and Twitter profile datasets are utilized to conduct the experiments on the proposed classification model. The study incorporates both balanced and imbalanced datasets to evaluate the performance and efficiency of the proposed model. These datasets can be accessed on Kaggle, a well-known online community for data science and ML researchers that provides a wide range of data-related resources and tools for performing research. (https://www.kaggle.com/datasets).

From March 15th to March 19th, 2019, a web crawler acquired the Instagram fake spammer dataset, which is evaluated for balanced class. There are 696 instances in this dataset, 348 of which are real accounts and 348 of which are fake, consisting of 12 profile-related features. Second, the Facebook spam dataset, which includes public profiles collected via the Facebook API and Facebook Graph API, is used for the analysis, which is also available in Kaggle. It has 600 profiles, 500 of which are genuine and 100 of which are considered spam profiles, and it includes 15 features. Third, the Twitter spam dataset generated by Chen et al. [28], which consists of 1 lakh profiles, 5,000 of which are malicious and 12 features, is used for the analysis [28]. The dataset is made available to researchers studying spam detection, defining tweets with malicious URLs as spam (http://nsclab.org/nsclab/resources/). The three datasets used in the study are summarized in Table 3.

Table 3 Dataset description

S.No.	Dataset	OSN	#Features	#Samples (#Real:#Fake)	Type
1	Facebook spam	Facebook	15	600 (500:100)	Imbalanced
2	Twitter spam	Twitter	12	4929 (3474:1455)	Imbalanced
3	Instagram fake spammer	Instagram	12	696 (348:348)	Balanced

4.2. Evaluation metrics

To evaluate the performance of the proposed model using various classifiers, accuracy, precision, recall, and F1-score are most frequently used [5]. However, the overall performance of the classification model cannot be measured only with these metrics, especially for an imbalanced dataset. Thus, various other metrics, such as true positive rate (TPR), false positive rate (FPR), G-mean, and kappa statistics, are employed additionally to measure the effectiveness of the proposed model. More specifically, TPR measures the probability of actual positive samples that are correctly predicted by a model. FPR measures the probability of actual negative samples that are misclassified by a model. Accuracy is the percentage of correct predictions among all predictions. Precision is the ratio of accurate positive predictions to all of the model's positive predictions, and recall is the ratio of positive predictions made out of all positive instances in the dataset. The F-measure is a balanced measure that computes the harmonic mean of precision and recall. The G-mean measures inductive bias via a positive/negative precision ratio, and the classifier performs better in majority and minority classes with a larger G-mean. Kappa, which measures error reduction between classification and entirely random classification, is a useful metric for unbalanced datasets.

4.3. Result analysis

Initially, the performance of the proposed ensemble (PE) was assessed using the Facebook spam dataset. The results were compared with the eight individual conventional classifiers such as AdaBoost, LR, DT, KNN, RF, NB, XGBoost, and SVM. These algorithms were selected due to their extensive use in previous research, in which default parameters were utilized for ease of use. These algorithms were assessed with the original dataset, with SMOTE and SMOTE with cost-sensitive learning using accuracy (Acc.), precision (Pre.), recall (Rec.), and F1-score (F1) as the performance metrics. A grid search based on 10-fold cross-validation was used to find the best parameters for all of the base classifiers of the proposed model. This improves performance and gives more accurate classification results. In general, the k-fold method is a technique used to evaluate ML models on a small data sample, requiring only one parameter, k, which represents the number of categories to divide the instances.

For the neural network model, three hidden layers with 16 hidden units in the first hidden layer, 8 hidden units in the second layer, and 4 hidden units in the last layer were included with a sigmoid function as the activation function. To ensure the validity of the PE, each experiment was performed 10 times, and the average values were employed for comparisons of results [5]. The results are presented in Table 4.

Table 4 Classification results of different ML on the Facebook spam dataset

Classifiers	Original dataset				With SMOTE				SMOTE with cost-sensitive			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
AB	0.88	0.88	0.88	0.88	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92
LR	0.86	0.86	0.86	0.86	0.91	0.91	0.91	0.91	0.93	0.93	0.93	0.93
KNN	0.89	0.89	0.90	0.86	0.91	0.91	0.92	0.91	0.92	0.92	0.92	0.91
RF	0.80	0.81	0.80	0.80	0.91	0.91	0.91	0.91	0.92	0.93	0.92	0.92
XGBoost	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92	0.93	0.92	0.93	0.93
DT	0.86	0.87	0.86	0.86	0.92	0.92	0.92	0.92	0.92	0.91	0.92	0.91
SVM	0.90	0.90	0.90	0.90	*0.94*	*0.94*	*0.94*	*0.94*	0.94	0.94	0.94	0.94
NB	0.88	0.88	0.88	0.88	*0.93*	*0.94*	*0.94*	*0.94*	0.94	0.94	0.94	0.94
PE	*0.92*	*0.92*	*0.92*	*0.92*	*0.94*	*0.94*	*0.94*	*0.94*	*0.95*	*0.96*	*0.95*	*0.95*

The obtained results indicate the efficiency of applying SMOTE and SMOTE with cost-sensitive learning on the dataset having imbalanced class distribution. With the original dataset, most classifiers performed well but had poor recall due to class imbalance. Thus, the trade-offs between precision and recall were seen in lower F1-scores. However, all classifiers improve when SMOTE balances the dataset. This indicates that SMOTE reduces class imbalance by improving accuracy, precision, recall, and F1-scores. The performance of all the classifiers used for the evaluation was improved even more by the incorporation of cost-sensitive learning strategies. The PE technique has a high F1-score, suggesting superior prediction accuracy than individual classifiers. Thus, the combination of SMOTE and cost-sensitive learning greatly improves classifier performance, especially in ensemble models, leading to a better classification rate.

Next, the Instagram fake spammer dataset was used for investigation on the accuracy of the proposed model, and the results were compared with the seven commonly used machine-learning methods, including LR, NB, SVM, KNN, neural network, RF, and XGBoost. This analysis was made to evaluate the performance of the proposed model on a balanced dataset, so applying SMOTE in the first phase was omitted.

Table 5 Comparison of accuracy of classifiers with Instagram fake spammer dataset

Classifier	Accuracy with (5 folds)	Accuracy with the split dataset
LR	95.70%	96.70%
NB	96.30%	97.20%
SVM	96.30%	97.80%
KNN	93.00%	94.40%
Neural network	93.00%	94.40%
RF	96.30%	96.10%
XGBoost	96.50%	96.60%
Proposed model	96.90%	98.20%

The evaluation of the results was attributed to two distinct approaches: k-fold (5) and dividing the dataset into 0.70 for training and 0.30 for testing. The results obtained for the Instagram fake spammer dataset are listed in Table 5. The results show that all the classifiers, in general, perform well in both contexts. However, all the classifiers except SVM offer improved performance with a split dataset. The classifiers such as NB, SVMs, RF, XGBoost, and the proposed model offered improved accuracy with a 5-fold valuation. On the other hand, classifiers such as NB, SVMs, and the proposed model showed superior performance compared to other classifiers with a split dataset. Moreover, with an accuracy of 96.90% in cross-validation and 98.20% on the split dataset, the proposed model consistently ranked toward the top. These findings indicate that the suggested

model has the potential for fake profile detection even with a balanced dataset. The results indicate that the proposed model achieves better results than other conventional classifiers not only for an imbalanced dataset (Facebook spam) but also for a balanced dataset (Instagram spam).

Subsequently, with the Twitter spam dataset collected by Chen et al. [28], the experiments were performed. The proposed method was tested against seven conventional classifiers, such as SVM, NB, DT, KNN, RF, LR, and XGBoost. The results were also compared with advanced classifiers such as cost-sensitive learning as improved deep neural networks (CSDNN), AdaCost [29], MetaCost [26], weight-selection strategy for deep neural networks (WSNN) [30], ensemble learning (EL) with majority voting, and cost-sensitive stacking (CSS) [5]. As each performance metric indicates a distinct feature of model performance, the various models are evaluated using TPR, FPR, precision, F1-score, G-mean, and Kappa. Table 6 displays the classification results of the Twitter spam dataset.

Table 6 Classification results of various classifiers with the Twitter spam dataset

Type	Method	TPR	FPR	Precision	F1-score	G-mean	Kappa
Conventional classifiers	SVM	0.10	0.01	0.77	0.18	0.31	0.16
	NB	0.91	0.81	0.10	0.18	0.41	0.02
	DT	0.57	0.05	0.53	0.55	0.74	0.50
	KNN	0.47	0.02	0.65	0.55	0.68	0.51
	RF	0.70	0.09	0.71	0.69	0.71	0.61
	LR	0.62	0.07	0.64	0.64	0.53	0.49
	XGBoost	0.69	0.08	0.69	0.71	0.81	0.60
Advanced classifiers	CSDNN [5]	0.54	0.05	0.52	0.53	0.71	0.48
	AdaCost [29]	0.67	0.09	0.41	0.51	0.78	0.45
	MetaCost [26]	0.69	0.10	0.40	0.51	0.77	0.45
	WSNN [30]	0.63	0.08	0.44	0.52	0.76	0.46
	EL [5]	0.69	0.04	0.63	0.66	0.81	0.62
	CSS [5]	0.70	0.03	0.70	0.70	0.82	0.67
	Proposed model	0.81	0.04	0.81	0.82	0.85	0.79

The obtained results indicate that the class imbalance greatly affects the performance of the classifiers. For instance, the lower TPR and G-mean of SVM indicate that it classifies the fake profiles as genuine profiles. On the other hand, though the TPR is high for NB, the FPR is high, causing lower kappa values, which indicates that genuine profiles are classified as fake. Similarly, classifiers such as DT and KNN provide average performance, while LR and XGBoost show above-average performance. Thus, among all these traditional classifiers, the RF produced improved results. However, the proposed method has a high TPR of 0.81, indicating the effective classification of positive instances (fake profiles), and a lower 0.04 FPR, indicating the effective classification of the majority class (genuine profiles). Moreover, the higher precision of 0.81 and F-measure of 0.82 demonstrate its efficacy in achieving a balance between precision and recall.

The results also indicate that the advanced classifiers show improved performance compared to conventional classifiers. Specifically, MetaCost offered improved TPR compared to CSDNN, AdaCost, and WSNN classifiers. However, its FPR is also high, indicating a higher misclassification of genuine profiles. Similarly, the performance of ensemble learning and CSS offered improved results with high TPR, G-mean, and Kappa values; however, these results are lower than the proposed model, which has an increased TPR (0.81), FPR (0.04), precision (0.81), and F1-score (0.82), indicating superior performance. The proposed model has a G-mean of 0.85, a metric that measures the balance between sensitivity and specificity, indicating that it can provide a good compromise between true positives and true negatives. Also, Kappa, a measure of agreement beyond random variation, shows a substantial value (0.79). The results show that the suggested method works well, with a high TPR, precision, F1-score, G-mean, and Kappa. It has the potential to identify fake profiles more accurately while keeping the FPR low. The analysis of various classifiers for the Twitter spam dataset is visualized in Fig. 3.

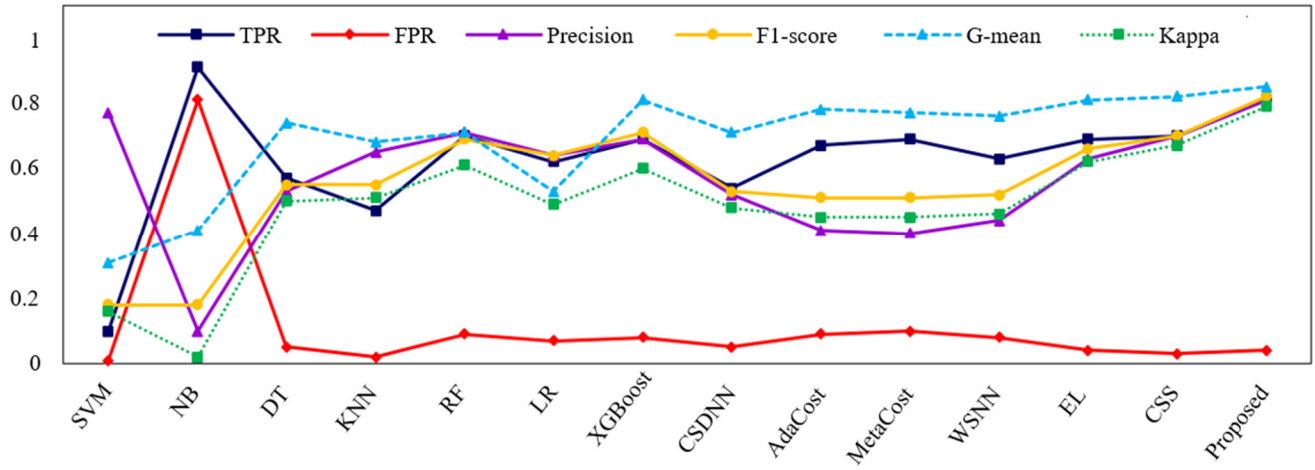

Fig. 3 Performance analysis of various classifiers for the Twitter spam dataset

Finally, the performance of the proposed model is also compared with the various existing models. Since there is no single dataset is employed by the existing works, it is difficult to compare the results of the various models. However, the results of a comparison study on the efficacy of existing methods for identifying spam or fake accounts on various OSNs using their synthesized datasets are displayed in Table 7.

Table 7 Classification results of different state-of-the-art classifiers

Existing work	OSN assessed	Accuracy (in %)
Albayati and Altamimi, 2019 [18]	Facebook	95.72
Elyusufi et al., 2019 [9]	Facebook	99.30
Dey et al. 2019 [20]	Instagram	90.80
Akyon and Kalfaoglu, 2019 [14]	Instagram	86.00
Liang et al., 2020 [11]	Sina Weibo	90.60
Sahoo and Gupta, 2020 [19]	Facebook	99.52
Purba et al., 2020 [8]	Instagram	91.76
Kaushik et al., 2022 [21]	Instagram	91.00
Mughaid et al., 2023 [12]	Facebook	97.80
	Instagram	94.70
Proposed model	Instagram	95.18
	Facebook	98.20

The results indicate that the existing works have higher performance with an accuracy greater than 90%. However, the majority of existing models, as indicated in prior literature [8, 11, 14, 20], lack accuracy. Furthermore, despite the higher accuracy of 99.30% for the method proposed in Elyusufi et al. [9], important details such as the feature selection process are not included in the study. Also, the models presented by Mughaid et al. [12] and Akyon and Kalfaoglu [14] are limited in sample size and lack detailed analysis with large datasets. Further, the model in Sahoo and Gupta [19] achieved a 99.52% accuracy rate, but it requires further validation using other standard datasets. Also, some models like those proposed by Sahoo and Gupta [19] and Dey et al. [20] lack comparison analysis with advanced classifiers, while the model presented by Kaushik et al. [21] struggles with short profiles. While the performance of the proposed model is comparatively lower than that of Elyusufi et al. [9] and Sahoo and Gupta [19], the overall result indicates that the proposed model outperforms the majority of contemporary approaches to fake profile detection.

Upon extensive evaluation and detailed analysis, the study successfully realized the formulated working hypothesis. The multistage stacked ensemble model, integrating chi-squared feature-class association-based feature selection and cost-sensitive learning, demonstrated a remarkable improvement over conventional and advanced classifiers in fake profile detection across various OSNs. The achieved accuracy and precision rates also ensure the effectiveness of the proposed approach in overcoming the challenges posed by the imbalanced datasets.

5. Conclusions

Owing to the significance of fake profile detection, this study introduces a novel multistage stacked ensemble classification model to enhance fake profile identification efficiency in OSN by overcoming unbalanced dataset challenges. This model has three stages. The first stage uses a chi-squared feature-class association model to choose features. This is followed by stacked ensemble classification with cost-sensitive learning and meta-cost-based neural networks for effective classification.

The proposed model is evaluated using three publicly available datasets and assessed with a wide range of performance indicators. Using cost-sensitive learning and ensemble learning techniques, the model has demonstrated superior performance in detecting fake profiles across several social media sites, including Facebook spam (an imbalanced dataset) and Instagram (a balanced dataset), with remarkable accuracy rates of 95% and 98.02%, respectively. These results also indicate that the model is not only suitable for imbalanced datasets but also for balanced datasets. The results of the proposed model with Twitter spam datasets also show the superior performance of the proposed method with 91% precision, compared with the conventional and advanced classifiers, demonstrating more precise identification of fake profiles while maintaining low FPR. The comparison of the results with the state-of-the-art classifiers also serves as critical validation, confirming the realization of the formulated hypothesis. The findings of the study state that the proposed model significantly improves the accuracy of fake profile detection, reducing the challenges posed by the imbalanced class problem.

Though the proposed multi-stage stacked ensemble classification model demonstrates superior performance in detecting fake profiles, the efficiency of the model may vary depending on the nature of the profiles assessed. Thus, it is essential to identify the effectiveness of the model by varying the dataset features and the nature of fake profiles. So, future enhancements could include various other feature engineering techniques such as NLP, image and network analysis for text and pictures, and behavior analysis for improving the accuracy of fake profile classification. Moreover, the widely used method for detecting fake profiles, such as similarity analysis, which assesses the similarity between the profiles along with fuzzy decision-making, can be used to assess the efficiency of the proposed model. Further, deep learning techniques like CNNs and recurrent neural networks (RNNs) could be used in extracting features and classification, expanding the ability of the model to detect fake profiles across multiple social media platforms.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] N. Thakur, “Social Media Mining and Analysis: A Brief Review of Recent Challenges,” Information, vol. 14, no. 9, article no. 484, September 2023.

[2] P. Wanda, “RunMax: Fake Profile Classification Using Novel Nonlinear Activation in CNN,” Social Network Analysis and Mining, vol. 12, no. 1, article no. 158, December 2022.

[3] R. Kaur, S. Singh, and H. Kumar, “Rise of Spam and Compromised Accounts in Online Social Networks: A State-of-the-Art Review of Different Combating Approaches,” Journal of Network and Computer Applications, vol. 112, pp. 53-88, June 2018.

[4] B. Drury, S. M. Drury, M. A. Rahman, and I. Ullah, “A Social Network of Crime: A Review of the Use of Social Networks for Crime and the Detection of Crime,” Online Social Networks and Media, vol. 30, article no. 100211, July 2022.

[5] C. Zhao, Y. Xin, X. Li, Y. Yang, and Y. Chen, “A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data,” Applied Sciences, vol. 10, no. 3, article no. 936, February 2020.

[6] A. Hassan, A. G. I. Alhalangy, and F. Alzahrani, “Fake Accounts Identification in Mobile Communication Networks Based on Machine Learning,” International Journal of Interactive Mobile Technologies, vol. 17, no. 04, pp. 64-74, February 2023.

[7] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, “Feature Selection Using an Improved Chi-Square for Arabic Text Classification,” Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 2, pp. 225-231, February 2020.

[8] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "Classification of Instagram Fake Users Using Supervised Machine Learning Algorithms," International Journal of Electrical and Computer Engineering, vol. 10, no. 3, pp. 2763-2772, June 2020.
[9] Y. Elyusufi, Z. Elyusufi, and M. H. A. Kbir, "Social Networks Fake Profiles Detection Based on Account Setting and Activity," Proceedings of the 4th International Conference on Smart City Applications, pp. 1-5, October 2019.
[10] M. S. Karakaşlı, M. A. Aydin, S. Yarkan, and A. Boyaci, "Dynamic Feature Selection for Spam Detection in Twitter," International Telecommunications Conference: Lecture Notes in Electrical Engineering, vol. 504, pp. 239-250, 2019.
[11] J. Liang, P. Jin, L. Mu, and J. Zhao, "Detecting Spammers from Hot Events on Microblog Platforms: An Experimental Study," The 32nd International Conference on Software Engineering and Knowledge Engineering, pp. 445-450, July 2020.
[12] A. Mughaid, I. Obeidat, S. AlZu'bi, E. A. Elsoud, A. Alnajjar, A. R. Alsoud, et al., "A Novel Machine Learning and Face Recognition Technique for Fake Accounts Detection System on Cyber Social Networks," Multimedia Tools and Applications, vol. 82, no. 17, pp. 26353-26378, July 2023.
[13] A. Sallah, E. A. A. Alaoui, and S. Agoujil, "Interpretability Based Approach to Detect Fake Profiles in Instagram," International Conference on Networking, Intelligent Systems and Security: Lecture Notes on Data Engineering and Communications Technologies, vol. 147, pp. 306-314, 2022.
[14] F. C. Akyon and M. E. Kalfaoglu, "Instagram Fake and Automated Account Detection," Innovations in Intelligent Systems and Applications Conference, pp. 1-7, October-November 2019.
[15] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine Learning-Based Social Media Bot Detection: A Comprehensive Literature Review," Social Network Analysis and Mining, vol. 13, no. 1, article no. 20, December 2023.
[16] A. Sallah, E. A. Abdellaoui Alaoui, S. Agoujil, and A. Nayyar, "Machine Learning Interpretability to Detect Fake Accounts in Instagram," International Journal of Information Security and Privacy, vol. 16, no. 1, pp. 1-25, 2022.
[17] I. Aydin, M. Sevi, and M. U. Salur, "Detection of Fake Twitter Accounts with Machine Learning Algorithms," Proceedings of International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-4, September 2018.
[18] M. B. Albayati and A. M. Altamimi, "Identifying Fake Facebook Profiles Using Data Mining Techniques," Journal of ICT Research and Applications, vol. 13, no. 2, pp. 107-117, September 2019.
[19] S. R. Sahoo and B. B. Gupta, "Fake Profile Detection in Multimedia Big Data on Online Social Networks," International Journal of Information and Computer Security, vol. 12, no. 2-3, pp. 303-331, 2020.
[20] A. Dey, H. Reddy, M. Dey, and N. Sinha, "Detection of Fake Accounts in Instagram Using Machine Learning," AIRCC's International Journal of Computer Science and Information Technology, vol. 11, no. 5, pp. 83-90, October 2019.
[21] K. Kaushik, A. Bhardwaj, M. Kumar, S. K. Gupta, and A. Gupta, "A Novel Machine Learning-Based Framework for Detecting Fake Instagram Profiles," Concurrency and Computation: Practice and Experience, vol. 34, no. 28, article no. e7349, December 2022.
[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
[23] C. V. Swetha, S. Shaji, and B. M. Sundaram, "Feature Selection Using Chi-Squared Feature-Class Association Model for Fake Profile Detection in Online Social Networks," The 3rd International Conference on Advanced Computing and Intelligent Technologies, article no. 24, December 2023.
[24] J. Yan and S. Han, "Classifying Imbalanced Data Sets by a Novel Re-Sample and Cost-Sensitive Stacked Generalization Method," Mathematical Problems in Engineering, vol. 2018, article no. 5036710, January 2018.
[25] P. Sterner, D. Goretzko, and F. Pargent, "Everything Has Its Price: Foundations of Cost-Sensitive Machine Learning and Its Application in Psychology," Psychological Methods, in press. https://doi.org/10.1037/met0000586
[26] P. Domingos, "MetaCost: A General Method for Making Classifiers Cost-Sensitive," Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155-164, August 1999.
[27] N. Ghatasheh, H. Faris, I. AlTaharwa, Y. Harb, and A. Harb, "Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks," Applied Sciences, vol. 10, no. 7, article no. 2581, April 2020.
[28] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, et al., "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection," IEEE Transactions on Computational Social Systems, vol. 2, no. 3, pp. 65-76, September 2015.
[29] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification Cost-Sensitive Boosting," Proceedings of the Sixteenth International Conference on Machine Learning, pp. 97-105, June 1999.
[30] A. Sze-To and A. K. C. Wong, "A Weight-selection Strategy on Training Deep Neural Networks for Imbalanced Classification," International Conference Image Analysis and Recognition: Lecture Notes in Computer Science, vol. 10317, pp. 3-10, 2017.