Facets of Fakes in Cyberspace: Machine and Ensemble Learning-Based Decisions and Detections

Ram Chatterjee[1*], Mrinal Pandey[1], Hardeo Kumar Thakur[2], Anand Gupta[3]
[1]Department of Computer Science & Technology, Manav Rachna University, Sector – 43, Aravalli Hills, Delhi – Surajkund Road, Faridabad, Haryana, India
[2]School of Computer Science Engineering and Technology, Bennett University, Plot Nos 8, 11, TechZone 2, Greater Noida, Uttar Pradesh, India
[3]Department of Computer Science and Engineering, Netaji Subhas University of Technology (NSUT) Sec-3, Dwarka New Delhi 110078, India
E-mail: ram@mru.edu.in, mrinalpandey@mru.edu.in, hardeo.thakur@bennett.edu.in, anand.gupta@nsit.ac.in
[*]Corresponding author

Keywords: information credibility, opinion spams, generalized additive2 classifier, elastic-net classifier, logloss, roc-auc

Received: September 1, 2024

Fake online reviews hinder internet marketing efforts to build businesses and brands in a competitive market with changing consumer expectations. This helps brands attract clients, making fake online reviews hard to uncover. Hence, fake reviews and websites are extensively examined. AI models like the Generalized Additive2 Model (GA2M) and its ensemble with the Elastic-net Classifier model have been studied using Log-Loss metric. This research, analysis, and depiction help demarcate bogus hotel reviews and websites from genuine entities. The paper uses ML classifiers (Decision Tree, Logistic Regression, Naïve Bayes) and ensemble models (Random Forest, Gradient Boosting) to identify legitimate websites using binary classification. This article compares ML classifiers and ensemble models by accuracy, precision, recall, f1-score, and ROC-AUC to evaluate their pros and downsides. Elastic-Net Classifier (L2 / Binomial Deviance) with score of 0.2879 outperformed GA2M model by 0.66% in LogLoss holdout score on Hotel dataset. LogLoss predicts values better than ROC-AUC due to its closer proximity to predicting actual values. Elastic-Net Classifier (L2 / Binomial Deviance) surpassed GA2M in F1 score, precision, and accuracy by 0.4%, 1.84%, and 0.63%. Ensemble techniques outperform ML classifiers in the Fraudulent and Legitimate Online Shops dataset with ROC-AUC scores of 0.71%, 1.73%, 0.76%, 1.10%, and 0.63% using 50% to 90% training datasets and 50% to 10% holdout datasets.

Povzetek: Raziskava je uporabila strojno in ansambelsko učenje (Elastic-Net, GA2M, Random Forest) za odkrivanje lažnih spletnih recenzij hotelov in goljufivih spletnih trgovin. Ugotovljeno je, da so ansambelski modeli bistveno boljši od posameznih klasifikatorjev pri prepoznavanju spletnih goljufij.

1 Introduction

The number of fakes in internet is growing in a stubborn, dormant, and questionable way. Spanning from fake news, bogus reviews, counterfeit websites, fictitious images and videos, the rapid dissemination of fakes [1] is surpassing, dominating, and overriding the online fake facade. The spread of misinformation and disinformation [2] presents significant challenges for individuals, societies, and nations alike. It affects public discourse, trust in information sources, and democratic processes. Therefore, this paper explores the facets of fakes in cyberspace by experimenting on datasets attributed to fake reviews and fake websites, implicating machine learning and ensemble learning directives of the well-known models, to decipher and demonstrate their deliberations and detections as inherent and coherent ability of these models, to distinguish between genuine and fakes.

The spread of inaccurate information in the online environment has serious implications in various fields: [3]

- Misinformation and disinformation:

Falsified content plays a role in disseminating misinformation, which refers to untrue or erroneous information shared without the intention to cause harm, and disinformation, which implicates deliberate propagation of false information to cuckold or manipulate. Both categories of fraudulent content erode the trustworthiness of sources of information and warp public understanding of reality [4].

- Social Disagreement and divide: The deceptive messages influenced by bogus news and wrought stories often take benefit of the current communal rifts, causing them to widen the gap between people and hinder the constructive dialogs which reinforce pre-existing misbeliefs and biases.
- Loss of Faith: The online sources, social networking sites, and conventional media channels are the centers of prevalent and ever-growing misleading

information which in turn compromises with the faith of people on them. As people are exposed to a growing amount of deceptive or untrue material, they tend to be more doubtful and less critical, resulting in a loss of trust.

- Intimidations to Democracy: In democratic nations, the use of false information to influence public opinion presents a notable risk to the fairness of elections and democratic systems. Fabricated content can impact how voters act, create uncertainty about election results, and weaken the credibility of democratic establishments.
- Brand promotion and demotion affecting product/service marketability: Product/service reviews posted online on e-commerce sites greatly influence brand/product marketability impacting customers' decisions on future purchases, impacted by counterfeit reviews, entailed by the diaspora of opinion spams, opinion spammers and collusive opinion spammers convoluting the phony reviews.

2 Information credibility - concerns

The evaluation of information credibility is a challenge and a prospect due to the fine-grained, significant, perceptive, and interrelated analysis and appraisal of product reviews and fraudulent websites using current approaches. This is due to the following reasons: [5]:

- Consumers' confidence in the brand and its products is bolstered by both positive and negative assessments, regardless of whether they are genuine or fabricated. They influence not only current purchase decisions but also future purchases and the overall consumer perception of the brand. Positive reviews enhance brand credibility and product promotion, while negative reviews, even if genuine, can have adverse effects. Additionally, fraudulent negative reviews can significantly damage consumer trust in the brand and its products, leading to reputational harm for the company with consequential losses in product marketing and business performance [6, 7].
- Customer feedback is a key factor influencing product rankings on the platform, which subsequently determines their purchasing decisions. The availability of huge variety of brands offering plenty of products to choose from augmented with their rankings influence the customers' decision on buying or rejecting the product. In addition to this, the product ranking algorithms are also swayed by product reviews. This confirms the significance of product reviews as a central aspect rendering product ranking that impacts customers' purchasing decisions. Owing to this reason product reviews are manipulated as bogus reviews by opinion spammers impacting its credibility and purchase potential leading to its promotion / demotion that in turn impresses the brand's survival in the competitive market.
- The threat and form of fake gets augmented with the existence and growth of deceptive shopping websites that replicates original shopping portals too closely to confuse customers to believe it as authentic webpage, which in turn let customers disclose their sensitive personal and financial credentials for compromise [8].

3 Literature review

3.1 The theoretical aspects

The academic inscription of literature survey mentions the emergence of fake online reviews in 2007 [9, 10] wherein linguistic features implicated in review text, behavioral features attributed to opinion spammers and product review characteristics were the strategies to ascertain bogus reviews. The progress further has led dissemination of fabricated reviews by group opinion spammers influenced by its well-paid option, and market demand of spammers swayed by brand promotion, profitability and competition survival. Consequently, consumers must critically assess reviews to distinguish between authentic and inauthentic ones when evaluating the credibility of products or services. [11].

The misleading review suggests fictitious writing about the product without any authentic firsthand experience. Furthermore, recent advancements in NLP have enabled the creation of false reviews on a large scale identical to genuine human-generated content. "Review spams" are attributed with their distribution on e-commerce platforms and social media outlets for the purpose of promoting products and undermining rival brands. The prevalence of "incentivized reviews" is also increasing, often resulting from brand countersignatures. While opinion spams may be authored by unidentified individuals, paid reviews are typically written by remunerated or sanctioned opinion promoters and may be identified. Therefore, in addition to addressing fake opinions directly, current research is also focused on understanding the behavioral tendencies of opinion promulgators as well as collective efforts to detect fake reviews [12, 13].

The propagation of fakes is just not limited to the fabricated product reviews but has also propagated in the form of fake websites that promote online sales. The zeitgeist juncture of retail businesses has initiated the process of digital transformation to offer their commodities and amenities online [14]. Both established and emerging companies are transitioning towards e-commerce platforms in order to connect with customers and showcase their product offerings [15]. These websites typically share a similar format to ensure easy access and user-friendliness for anyone interested in the brand or company. Opportunists exploit this uniformity by setting up fraudulent e-commerce stores that sell counterfeit products or engage in financial scams, duping unsuspecting customers out of their money.

3.2 The analysis of sham treatments

The digital era has resulted in an unparalleled increase in the sharing of information and online business transactions. Nonetheless, it has also caused a surge in spam, such as fabricated reviews and fraudulent websites [16], which can greatly erode user confidence and the authenticity of internet platforms. It is essential to detect these deceitful activities to uphold the dependability of digital environments. Recent advancements in machine learning and ensemble learning [17] have proven to be effective approaches for addressing the challenge of spam, enabling the development of sophisticated mechanisms for identifying and mitigating this issue.

Mechanisms for detecting spam [18] utilize a diverse set of attributes to identify and mitigate deceptive behaviors. These characteristics can be broadly categorized as content-related, behavioral, and metadata-based traits. In the context of academic texts, content-based features are frequently employed to analyze reviews or website content for signs of spam. Such features include lexical aspects like word and character count, as well as the presence of specific words or phrases that may indicate fraudulent reviews through irregular patterns, excessive promotional language, or repetitive expressions. Also, syntactic characteristics, such as sentence structure, grammar, and punctuation, are examined to identify poor language usage and unconventional sentence construction typically associated with spam content. Additionally, semantic features focus on evaluating the meaning and importance of the material by utilizing NLP (Natural Language Processing) methods [19, 20] to detect sentiment discrepancies within the context and tone of the content. The detection of fake reviews is achieved through the collaboration of NLP and ML. Initially, NLP is employed to process and convert text data into meaningful features, such as word frequencies, sentiment scores, and embeddings. Subsequently, these attributes are utilized in machine learning algorithms, including Naive Bayes, Decision Trees, and Logistic Regression, to identify reviews as either authentic or fraudulent. Random Forest and Gradient Boosting are examples of ensemble learning methods that incorporate multiple models to enhance accuracy and robustness by capturing intricate patterns in the data. By utilizing both behavioral patterns and linguistic cues, this synergy facilitates the more accurate identification of deceptive evaluations.

Behavioral Characteristics [21] analyze the behaviors and habits of individuals who publish reviews or establish websites. Important behavioral traits include the review habits that encompass both the frequency and timing of reviews. Dishonest reviewers may submit numerous reviews in a short time span or follow irregular posting schedules. The user conduct focuses on evaluation of user profiles, including the range of products reviewed and the reliability of ratings provided. Fraudulent reviewers frequently have new accounts, limited review track record, and prejudiced ratings. Augmenting it are the interaction styles that entails scrutinizing end users engagement with the website or other users. Unusual interaction patterns, such as an excessive number of clicks or rapid navigation, can suggest spam-related activity [8].

Additional information about the review or website is offered through metadata characteristics, such as IP address examination where detection of numerous reviews originating from a single IP address or geographical location may suggest fraudulent behavior [22]. Another metadata feature is time analysis attributing to uncommon timing patterns, such as multiple simultaneous reviews, may indicate automated spamming. Further the referral information metadata connoting to the assessment of the origin of website traffic can aid in identifying counterfeit websites, particularly if the traffic stems from questionable or unrelated sources [22].

The challenge lies in the idea that detecting fake reviews or dubious websites using AI-based methods is not feasible without human intervention to teach the AI algorithms, which are influenced by human biases introduced through dataset selection, data wrangling, selection of machine learning classification models, feature engineering techniques and adjusting hyper parameters for optimal performance. [23].

3.3 The dataset's delineation

The limited availability of fake review datasets is compounded by the difficulty of integrating each dataset into AI models for classification, due to differences in content and context.

Table 1 depicts the comparative analysis of the state-of-the-art methods used in the papers [6, 7, 24] proportional to the approach used in this research with an objective to emphasize that ensemble models produce better results as compared to individual ML classifiers used for experimentation purposes, as depicted in sections 5.1 and 5.2 respectively.

Figure 1 shows the gold standard dataset, utilized in the research, representing a well-balanced dataset that is suitable for testing AI models. A priori, textual data was preprocessed using standard preprocessing steps, which included lowercasing, removing punctuation, and tokenizing words. Additionally, feature extraction was performed to extract features such as unigrams, bigrams, part-of-speech tags, and psycholinguistic cues. These features are then numerically represented for machine learning models. Further, Exploratory Data Analysis involved the examination of linguistic patterns that were prevalent in dataset, such as the increased use of verbs and adverbs and the decreased use of nouns and concrete terms that are typical to deceptive reviews.

The Hotel dataset provided by [6, 7, 24] consists of 1600 reviews, encompassing both genuine and fraudulent feedback for 20 prominent Chicago hotels. These appraisals are bifurcated into two groups: eight hundred authentic and eight hundred fake reviews. The legitimate reviews comprise four hundred positive and four hundred negative evaluations.

In conclusion, the chosen dataset is made available for experimentation and research attributed with a well-

balanced distribution of optimistic and deleterious reviews, along with diversity in review length and hotels, which influences the diversity of review content and context. This makes it suitable for effective training to achieve promising results when analyzing the performance of different machine learning classifiers.

To consolidate and conclude on the performance of ML classifiers and ensemble learning models another dataset comprising illegitimate e-commerce site data alongside authentic e-commerce site data has been considered for the experimental purposes [25]. The dataset is well-balanced and comprises 1140 entries, with 579 representing fake (fraudulent) online shops and 561 representing real (legitimate) ones. Each entry includes the following attributes as depicted in Table 2 below with informative features (attributes potentially valuable for modeling) indicated in Figure 2.

Table 1: Comparative analysis with state-of-the-art methods [6, 7, 24]

Aspect	**Reference [7]**	**Reference [6]**	**This Paper**
Dataset Used	Deceptive Opinion Spam Corpus: The total number of hotel evaluations on TripAdvisor is 800, which includes 400 honest and 400 fraudulent reviews.	400 truthful and 400 bogus negative hotel reviews, comprising 800 reviews, comprise the Extended Deceptive Opinion Spam Corpus.	Extended Deceptive Opinion Spam Corpus [24]: **The total number of reviews is 1600,** with 400 truthful positive reviews from TripAdvisor, 400 phony positive reviews from Mechanical Turk, 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp, and 400 bogus negative reviews from Mechanical Turk.
Machine Learning Models	Support Vector Machines (SVM), Naïve Bayes, MaxEnt (Logistic Regression).	Unigram and bigram features are incorporated into SVM.	Generalized Additive2 Model (GA2M) and its ensemble with the Elastic-net Classifier model. **Ensemble models use many classifiers to improve performance.**
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score.	Accuracy, Precision, Recall, F1-Score.	Accuracy, Precision, Recall, F1-Score incorporating Log-Loss metric. **Log-Loss predicts values better than ROC-AUC due to its closer proximity to predicting actual values.**
Performance Results	Achieved up to 89.8% accuracy with SVM using unigrams and bigrams.	Achieved higher accuracy with bigram features compared to unigrams.	**GA2M achieved 92.5% recall. Enhanced outcomes, probably attributable to the application of ensemble methodologies and more extensive datasets.**
Techniques Used	The analysis of linguistic signals associated with deception; text classification using n-gram features.	Focused on negative sentiment reviews; utilized n-gram features and linguistic analysis.	Employed **ensemble learning techniques** to validate that it **provides better results than individual ML models.**
Limitations	Limited dataset size and domain specificity; potential overfitting; lack of generalizability to other domains.	Similar limitations as in [7] study; focus on negative reviews may not capture full spectrum of deceptive practices.	Limited dataset size and domain specificity; potential challenges in interpretability of ensemble models.

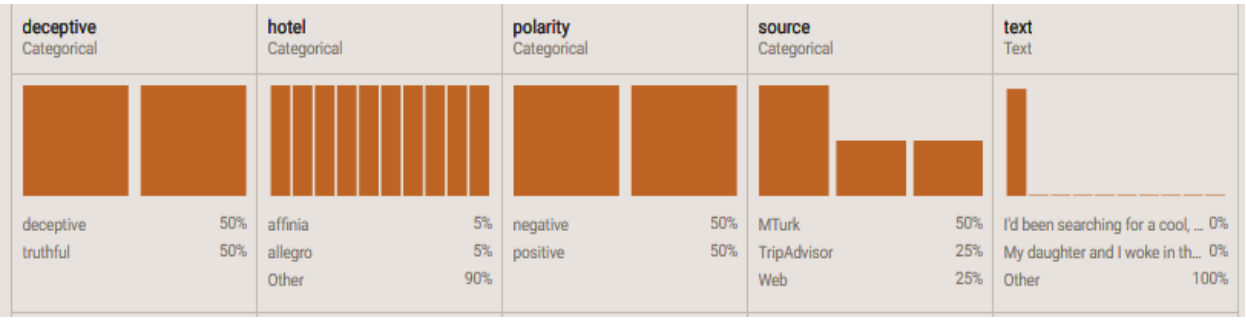


Figure 1: Gold standard hotel dataset preview with 5 informative features

Table 2: Fraudulent and legitimate online shops dataset attributes [25]

Sl. No.	Dataset Attribute
1.	Online shop's URL;
2.	Label - {legitimate, fraudulent};
3.	Domain length - Count of symbols in the host domain name;
4.	Top domain length - Count of symbols in the top domain name;
5.	Presence of prefix "www" in the active URL of the online shop, values {0 - no, 1 - yes};
6.	Number of digits in the URL;
7.	Number of letters in the URL;
8.	Number of dots (.) in the URL;
9.	Number of hyphens (-) in the URL;
10.	Presence of credit card payment, values {0 - no, 1 - yes};
11.	Presence of money back payment, including Apple Pay, PayPal, Google Pay, Alipay, Samsung Pay, and Amazon Pay, values {0 - no, 1 - yes};
12.	Presence of cash on delivery payment, values {0 - no, 1 - yes};
13.	Presence of the ability to use crypto currencies for payments, values {0 - no, 1 - yes};
14.	Presence of free contact emails, including Gmail, Hotmail, Outlook, Yahoo Mail, Zoho Mail, Proton Mail, iCloud Mail, GMX Mail, AOL Mail, mail.com, Yandex Mail, Mail2World, or Tutanota, values {0 – email address not found, 1 - free email address, 2 - domain email address, 3 – other email address};
15.	Presence of logo URL, values {0 - no, 1 - yes};
16.	SSL certificate issuer name;
17.	SSL certificate expire date;
18.	SSL certificate issuer organization name;
19.	SSL certificate issuer organization ID, values {1 - Cloudflare, Inc., 2 - Let's Encrypt, 3 - Sectigo Limited, 4 - cPanel, Inc., 5 - GoDaddy.com, Inc., 6 - Amazon, 7 - DigiCert, Inc., 8 - Global Sign nv-sa, 9 - Google Trust Services LLC, 10 - ZeroSSL, 11 - other organization};
20.	Indication of young domain, registered 400 days ago or later, values {0 - 'old' domain name, 1 - 'young' domain name, 2 - 'hidden'};
21.	Domain registration date;
22.	Presence of TrustPilot reviews, values {0 - no, 1 - yes};
23.	TrustPilot score, values - real number from 0 to 5 or -1 if no reviews are available;
24.	Presence of SiteJabber reviews, values {0 - no, 1 - yes};
25.	Presence in the standard Tranco list, values {0 - no, 1 - yes};
26.	Tranco List rank, values - integer number from 1 to 1000000 or -1 if domain is not listed in the Tranco list.

Figure 2: Well balanced fraudulent and legitimate online shops dataset preview with 23 informative features

4 The experimental elucidation

4.1 AI models viabilities on hotel dataset

This section illustrates the models utilized to anticipate the target category, specifically regarding whether a review is authentic or fabricated within the projected tactic outlined in Figure 3. This implicates investigating with several machine learning classifiers and leveraging ensemble learning to enhance results. In all conducted experiments assessing model performance, there is a consistent use of "stratified sampling" for dataset partitioning, followed by further subjecting each subgroup to simple random sampling to ensure that the holdout dataset accurately represents the data while preserving stratum percentages. The experiment has been performed using the 5-fold and 10-fold cross-validation methods to safeguard against overfitting due to limited data availability and validate model performance.

The validation process is designed to address overfitting by utilizing the "training-validation-holdout" method. This method allocates 64% and 72% of the dataset for model training, 16% and 8% for model validation (which effectively addresses overfitting through cross-validation), and 20% for the final model evaluation. Figure 4 below illustrates the significance of this holdout dataset in assessing the model's effectiveness. Utilization of 64/16/20 is justified when the model necessitates calibration, and when enhanced validation stability is desired, or when the model exhibits a tendency towards overfitting. However, utilization of 72/8/20 is justified when the model gains from more training data and the requirements for adjustment are little or superficial.

The experimentation was conducted using an ensemble of validated top models, such as the Generalized Additive2 Model (GA2M) [26] and the Elastic-Net Classifier (L2 / Binomial Deviance) [27]. These models were ranked based on their success in resolving the binary classification problem of distinguishing between genuine and fake hotel reviews. As illustrated in Figure 5, this is explicitly described in the blueprint of the model.

Figure 3: Suggested Approach

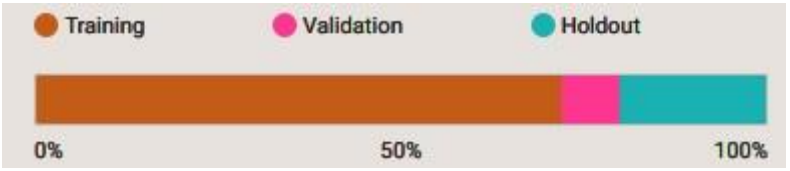


Figure 4: Segregation of the dataset into training, validation, and holdout sets utilizing 10-fold Cross-Validation

4.2 AI models viabilities on Fraudulent and Legitimate Online Shops Dataset

This section demonstrates the models used to predict the target category, particularly in determining if an e-commerce website is authentic or deceptive based on the "Label" attribute in the dataset following the proposed approach presented in Figure 6. This includes testing different machine learning classifiers and employing ensemble learning techniques to better the results. All evaluations on execution of the models were conducted implicating training datasets at 50%, 60%, 70%, 80% and 90% respectively with a 5-fold cross-validation assimilated to avoid overfitting, while the remaining percentage of the dataset was used as a holdout dataset respectively to validate the models' accomplishment on standard metrics.

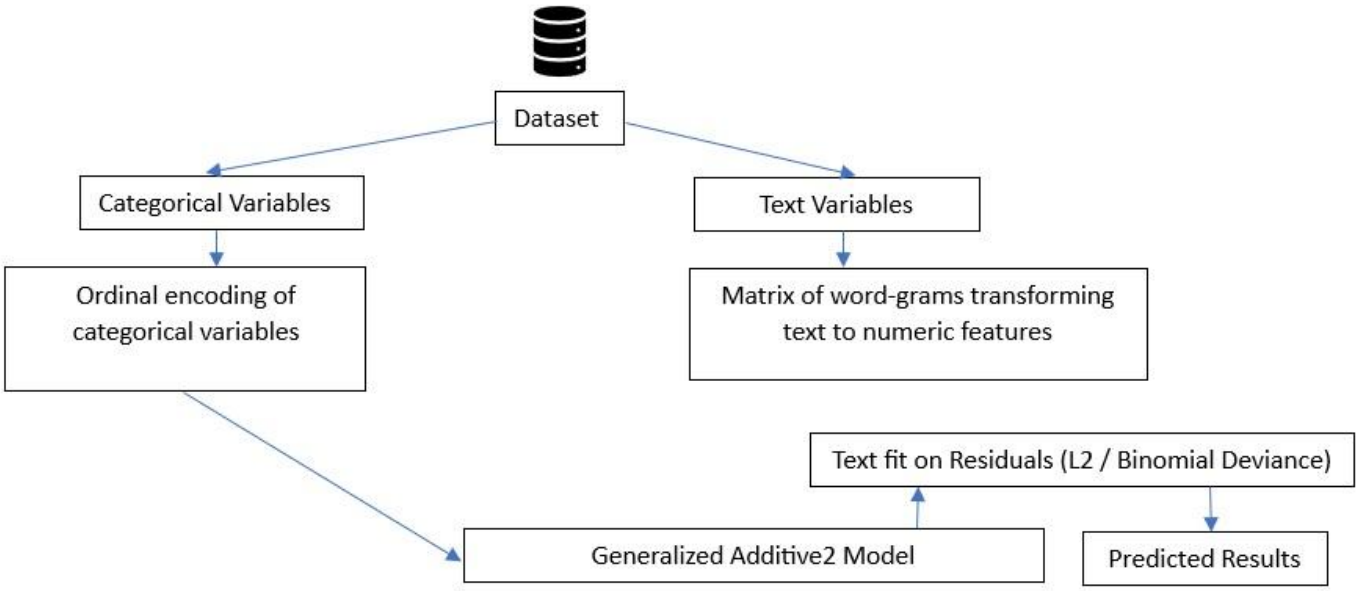


Figure 5: Ensemble of GA2M with Elastic-Net Classifier (L2 / Binomial Deviance)

Figure 6: Proposed methodology

5 The results' depiction

5.1 Experiment's revelations on hotel dataset

The results of the conducted model experiments mandate incorporating Log-Loss [28, 29] as the measurement for performance. The Log-Loss, commonly referred to as cross-entropy loss, reflects how closely the predicted probability matches to the actual genuine value and is considered superior to ROC-AUC. A lower log-loss value indicates better model achievement. The result of the experiment has been elaborated from Table 3 through Table 6 as depicted below.

Table 3: Data features for the AI models

Feature Name	Var Type	Unique	Missing	Target Leakage
deceptive	Categorical	2	0	N/A
hotel	Categorical	20	0	Low
polarity	Categorical	2	0	Low
text	Text	1277	0	N/A

Here, "Low" indicates that the experimented AI models take precautions from being overly optimistic during predictions.

Table 4: Implementation of the Elastic-Net Classifier (L2 / Binomial Deviance) Model on the Log-Loss Metric

Type of Scoring	Log Loss Metric Score
holdout	0.2879
validation	0.2745

Table 5: Achievement of alternative models' predictions on log-loss metric sorted by holdout score

Name of the Model	Validation Outcome	Holdout Outcome	% of Training Dataset
Elastic-Net Classifier (L2 / Binomial Deviance)	0.2745	**0.2879**	72.0
Generalized Additive2 Model (GA2M) with 10-fold CV	0.2536	0.2898	72.0
Elastic-Net Classifier (mixing alpha=0.5 / Binomial Deviance)	0.2743	0.2899	72.0
eXtreme Gradient Boosted Trees Classifier	0.2718	0.2903	72.0
Generalized Additive2 Model with 5-fold CV	0.2793	0.3075	64.0

Table 6: Key metric values of the models

Model Name	F1 Score	Recall	Precision	Accuracy
Elastic-Net Classifier (L2 / Binomial Deviance)	**0.8902**	0.9125	**0.869**	**0.8875**
Generalized Additive2 Model with 10-fold CV	0.8862	**0.925**	0.8506	0.8812
Elastic-Net Classifier (mixing alpha=0.5 / Binomial Deviance)	0.8855	0.9187	0.8547	0.8812
Generalized Additive2 Model with 5-fold CV	0.8822	0.9125	0.8538	0.8781
eXtreme Gradient Boosted Trees Classifier with Early Stopping	0.8779	0.9437	0.8207	0.8688

The Receiver Operating Characteristic curve illustrates the model's performance, metrics, and arrangement in the context of probability analysis. The AUC is depicted in relation to true positive and true negative rates with respect to the underlying data used in the study, as indicated by the shape of the curve and the Area Under the Curve. AUC summarizes performance as a single value by considering all potential thresholds for a binary classification problem [28, 30].

As illustrated in Figure 7 and Figure 8, the Generalized Additive2 Model (GA2M) demonstrates a superior ROC-AUC score and marginally higher Holdout score compared to Elastic-Net Classifier (L2 / Binomial Deviance), leading to the conclusion that GA2M is the more favorable model due to its "improved capability to discriminate" between matching and mismatching instances in the dataset. In amplification, it further exhibits (GA2M) as a better model with slightly improved "generalization performance" as indicated by its higher Holdout score when applied to unseen data [31]

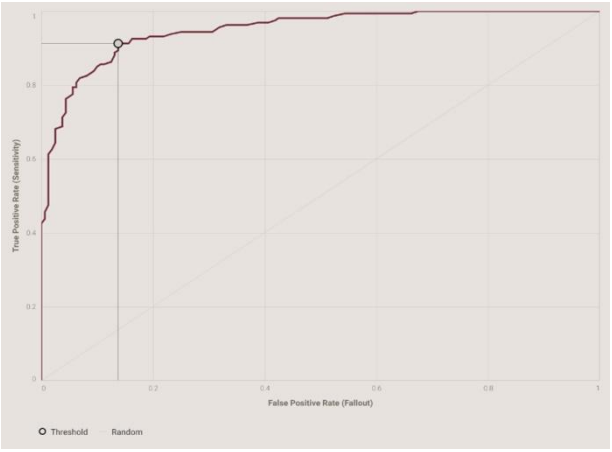


Figure 7: The ROC – AUC (0.9125) as calculated using the Holdout outcome (0.2879) of the Elastic-Net Classifier (L2 / Binomial Deviance).

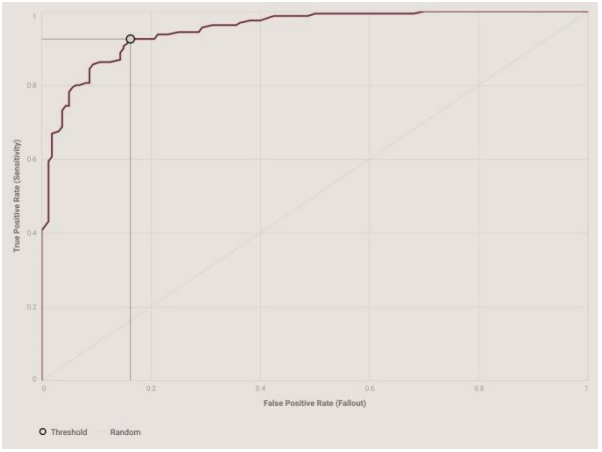


Figure 8: ROC – AUC (0.925) as calculated using the Holdout outcome (0.2898) of the Generalized Additive2 Model (GA2M) with a 10-fold CV

5.3 Experiment's revelations on fraudulent and legitimate online shops dataset

The results of the conducted model experiments with training dataset varying from 50% to 90% inclusive in training the varied number of ML classifiers and Ensemble Models with assessment on the test sets ranging from 10% to 50% of the dataset under consideration. The models' performance evaluation is promulgated on the basis of standard metrics [32] nominated for classification models viz. Accuracy, Precision, Recall, F1 score and ROC-AUC as indicated in Table 7 through Table 12 with the corresponding comparative analysis via graphs depicted in Figure 9.

Table 7: Accomplishment of AI models on training dataset and evaluation on accuracy metric on test data

Training Dataset %	Machine Learning Models			Ensemble Models	
	Decision Tree	Logistic Regression	Naïve Bayes	Random Forest	Gradient Boosting
50%	89.19%	**91.40%**	81.93%	**98.25%**	97.89%
60%	89.53%	**89.91%**	81.36%	98.25%	**98.46%**
70%	89.32%	91.81%	**94.15%**	**98.54%**	97.66%
80%	88.60%	91.23%	**95.18%**	**97.81%**	96.49%
90%	87.50%	85.96%	**95.61%**	**97.37%**	96.49%

Table 8: Accomplishment of AI models on training dataset and evaluation on precision metric on test data

Training Dataset %	Machine Learning Models			Ensemble Models	
	Decision Tree	Logistic Regression	Naïve Bayes	Random Forest	Gradient Boosting
50%	**93.41%**	89.63%	73.77%	**100.00%**	98.58%
60%	**93.50%**	89.79%	73.72%	**100.00%**	99.56%
70%	**95.04%**	92.40%	93.71%	**100.00%**	98.90%
80%	92.92%	92.44%	**95.08%**	**100.00%**	99.13%
90%	89.36%	88.33%	**95.24%**	**100.00%**	**100.00%**

Table 9: Accomplishment of AI models on training dataset and evaluation on recall metric on test data

Training Dataset %	Machine Learning Models			Ensemble Models	
	Decision Tree	Logistic Regression	Naïve Bayes	Random Forest	Gradient Boosting
50%	85.86%	93.71%	**99.30%**	96.50%	**97.20%**
60%	84.62%	90.56%	**98.71%**	96.58%	**97.44%**
70%	82.21%	91.33%	**94.80%**	**97.30%**	96.76%
80%	85.37%	90.91%	**95.87%**	**95.87%**	94.21%
90%	82.35%	85.48%	**96.77%**	**94.83%**	93.10%

Table 10: Accomplishment of AI models on training dataset and evaluation on F1-score metric on test data

Training Dataset %	Machine Learning Models			Ensemble Models	
	Decision Tree	Logistic Regression	Naïve Bayes	Random Forest	Gradient Boosting
50%	89.47%	**91.62%**	84.65%	**98.22%**	97.89%
60%	88.84%	**90.17%**	84.40%	98.26%	**98.49%**
70%	88.16%	91.86%	**94.25%**	**98.63%**	97.81%
80%	88.98%	91.67%	**95.47%**	**97.89%**	96.61%
90%	85.71%	86.89%	**96.00%**	**97.35%**	96.43%

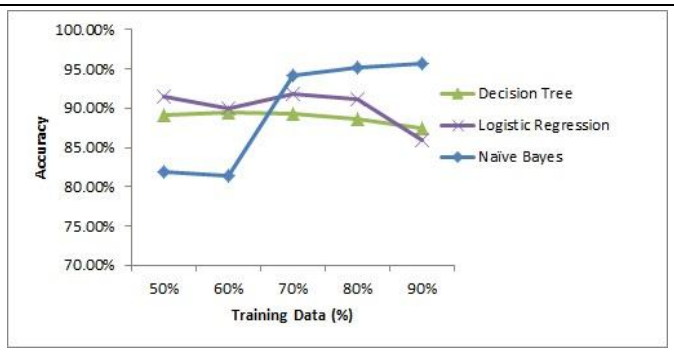
Table 11: Accomplishment of AI models on training dataset and evaluation on ROC-AUC metric on test data

Training Dataset %	Machine Learning Models			Ensemble Models	
	Decision Tree	Logistic Regression	Naïve Bayes	Random Forest	Gradient Boosting
50%	95.14%	96.19%	**99.11%**	99.78%	**99.81%**
60%	94.59%	96.58%	**98.14%**	99.77%	**99.84%**
70%	94.65%	96.90%	**99.13%**	**99.89%**	99.76%
80%	96.84%	97.29%	**98.84%**	**99.93%**	99.73%
90%	96.67%	95.32%	**99.32%**	99.86%	**99.94%**

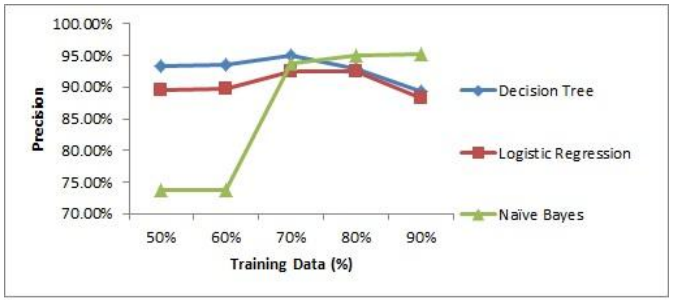
Table 12: Accomplishment of ensemble models on ML Classifiers on ROC-AUC metric indicating % increase in AUC

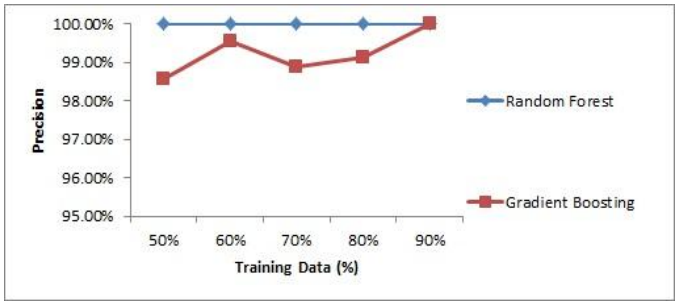
Training Dataset %	ROC-AUC of Ensemble Models	ROC-AUC of ML Classifiers	Percentage increase in AUC
50%	0.998140944 (GB)*	0.99109869 (NB)*	**0.71%**
60%	0.998421498 (GB)*	0.98144691 (NB)*	**1.73%**
70%	0.998863832 (RF)*	0.991312378 (NB)*	**0.76%**
80%	0.999304858 (RF)*	0.988414304 (NB)*	**1.10%**
90%	0.999384236 (GB)*	0.993176179 (NB)*	**0.63%**
*GB: Gradient Boosting	*RF: Random Forest	*NB: Naïve Bayes	

Graphs on ML Classifiers	Graphs on Ensemble Models

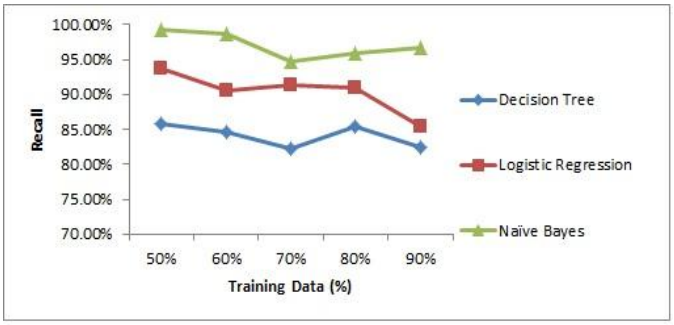


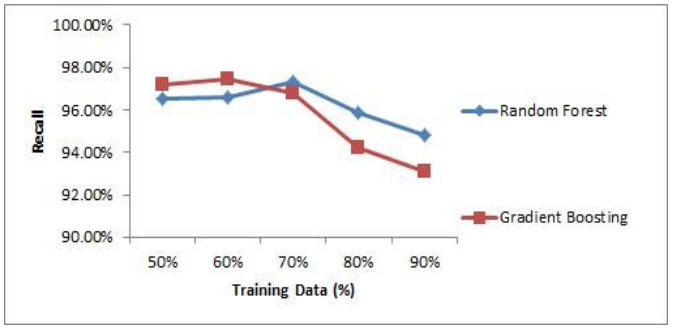
(a)



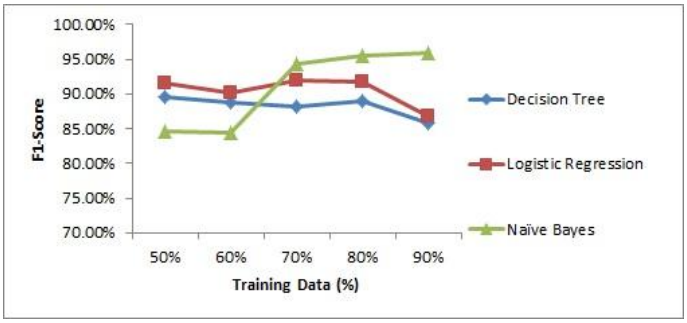


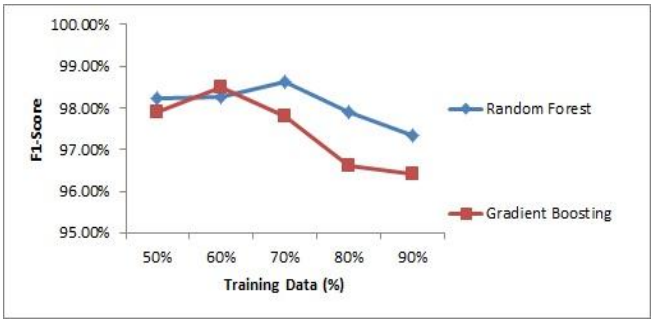
(b)



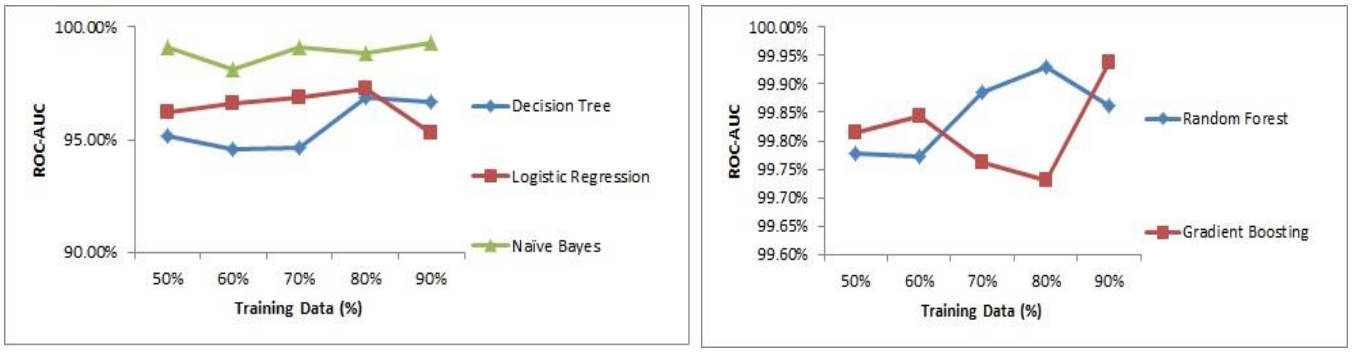


(c)





(d)



(e)

Figure 9: Performance comparison of ensemble models and ML classifiers in the areas of (a) accuracy (b) precision (c) recall (d) F1 score (e) AUC-ROC

6 The results' inspection

6.1 Machine learning classifiers vs. ensemble techniques – comparative analysis

- As depicted in section 5.2, Table 11, amongst the ML classifiers, Naïve Bayes has scored the best results on the standard metrics with ROC-AUC score ranging from 99.11% to 99.32%.
- Further observation on the performance of Ensemble Models depicted in section 5.2, Table 11 reveals far better score of standard metrics, wherein the Gradient Boosting Ensemble Model performs the best with 99.94% ROC-AUC score on 90% training data and 10% testing data.
- Comparing the outcomes of section 5.1 and 5.2 further consolidates that ensemble learning models are best in performance in context of both gold standard datasets and well balanced dataset (as indicated in Table 12 depicting percentage increase in AUC of ensemble models over ML classifiers), which manifests that ensemble learning directives like Random Forests and Gradient Boosting, frequently yield superior results compared to standalone ML classifiers by mitigating overfitting, enhancing stability, utilizing varied model viewpoints, minimizing errors, and boosting accuracy. Through amalgamating the strengths of several models, ensembles establish a more dependable and resilient predictive model.

6.2 The standard metrics – insight into the intricacies

The deliberations and directives on standard metrics [32] have been discussed below from the perspective of its application.

- Accuracy: The accuracy metric signifies the percentage of truthful forecasts (including true positives and true negatives) out of all predictions.

$$\text{Accuracy} = \frac{\text{True Positives (TP)}+\text{True Negatives (TN)}}{\text{Total number of instances}} \quad (1)$$

  - Pros: Easy to understand and implement.
  - Cons: Can be misleading with imbalanced datasets.

- Precision: Precision, which is also known as Positive Predictive Value, assesses the proportion of accurate positive identifications.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Positives (FP)}} \quad (2)$$

  - Pros: Important at instances where the price of false positives is high.
  - Cons: Does not cater for false negatives.

- Recall: Recall, which is also known as Sensitivity or True Positive Rate, quantifies the proportion of true positives that were accurately identified.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Negatives (FN)}} \quad (3)$$

  - Pros: Crucial at instances where the price of false negatives is high.
  - Cons: Does not consider false positives.

- F1-Score: The F1 Score is a unified evaluation of performance, representing the balanced harmonic mean of precision and recall.

$$F1\ Score = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

  - Pros: This proves helpful when imbalanced datasets are involved in experimentation.

- Cons: In relevance to F1 score other individual metrics may be more interpretable than this.

- ROC-AUC: A model's capacity to distinguish between classes is quantified by the ROC-AUC metric (Receiver Operating Characteristic - Area Under Curve). It accomplishes this by calculating the area under the ROC curve, which illustrates the compromise between the true positive rate and the false positive rate.
  - Pros: Evaluates the classification model's performance in relation to all decision thresholds.
  - Cons: Conceals the information regarding the precision of the calibration of the forecasted probability estimates.
    .

6.3 The LogLoss vs. AUC – discriminating features

- Probability Calibration:
  - As a matter of fact, rather than just considering the final classifications, the assertiveness of the model's individual predictions is the primary focus of LogLoss. It assesses the degree of agreement between the model's predicted probabilities and the actual results. which in turn is significant for the applications that require precise probability estimates [28], justifying its preferable use with the hotel dataset.
  - Comparatively, In contrast, the ROC-AUC metric [30] evaluates a model's ability to distinguish between distinct classes, irrespective of the precision of the predicted probabilities. As a result, a model may have a high ROC-AUC score even when it generates inaccurately calibrated probability estimates [32].

- Class Imbalance Sensitivity:
  - LogLoss is a metric that emphasizes the importance of accurately estimating all class probabilities, including for minority classes. This can make it more sensitive to class imbalances in the data [29]. But, as hotel dataset is a gold dataset, usage of Log-Loss metric suits the purpose of reflecting unbiased model calibration and accuracy.
  - The ROC-AUC metric is generally reliable even when the class distribution is highly unbalanced, as it assesses the ranking of predictions rather than their specific values. However, this metric may hide any issues the model has in dealing with minority classes [32].

- Interpretability and Directness:
  - LogLoss is a clear and easy-to-understand metric that is closely connected to the accuracy of probability predictions. Reduced LogLoss values indicate improved performance by capturing both the accuracy and confidence of predictions [29]. Log-Loss assesses the alignment of the model's probabilities with actual results connoting to its use with hotel dataset.
  - ROC-AUC evaluates the balance between the true positive rate (sensitivity) and the false positive rate (specificity) at multiple decision threshold values. It is valued for gauging the model's overall inequitable ability but may not directly offer actionable insights on probability calibration [32].

This concludes the fact that LogLoss is favored in situations that require accurately calibrated probability estimates and when addressing scenarios where the consequences of inaccurate predictions are significant. It provides a more comprehensive evaluation of the model's efficacy by imposing penalties for incorrect probability assignments whereas the ROC-AUC metric is valuable for assessing the model's overall capacity to differentiate among classes, particularly in imbalanced datasets. It gives less weightage to the specific predicted probabilities and emphasizes more on the ranking of predictions.

This confab signifies and justifies the reason of using LogLoss as mandated performance metric with gold dataset (the hotel dataset) depicted in the experimentation conducted in section 5.1 and implicating ROC-AUC as evaluation metric with well-balanced dataset (Fraudulent and Legitimate Online Shops) which isn't a gold standard dataset, depicted in experimentation conducted in section 5.2 as feasible, formal and recurrently used benchmarks for assessing models' performance in apt and befitting manner.

In order to mitigate the inherent bias of the LogLoss performance metric, it has been implemented in conjunction with the gold dataset. This approach involves stratified sampling, as well as 5-fold and 10-fold cross-validation, which facilitate the computation of the average LogLoss, thereby reducing variance and bias. Additionally, regularization (L2 penalty) has been promoted in order to prevent overconfident predictions. The same control on biases inherent to ROC-AUC with the well-balanced dataset has been exercised by complementing ROC-AUC with threshold-sensitive metrics such as Precision, Recall, F1 Score, and Accuracy at specific thresholds inherent to the experimentation conducted in section5.2.

7 Inference and imminent research

The virtual space for commodity appraisals has been heavily influenced by deceptive tactics, which have spread into various areas including service evaluations, sentiment analysis, star ratings, fake websites and other forms of deceitful behavior. This encompasses a wide range of

fraudulent activities from fake opinions to the individuals behind them who operate collusively.

Research in this rigid domain of identifying deceptive reviews has progressed rapidly, moving from analyzing linguistic features to examining behavioral and temporal patterns such as group spamming. Augmenting this implicates the analysis of stylistic and stylometric attributes of the appraisals. Furthermore, the application of machine learning models has been gradually improving, including ensemble approaches that encompass both traditional machine learning and deep learning model implementations.

The objective of the investigation is to test advanced AI models that use a combination of methods for better results in binary classification. It focuses on detecting fake/genuine reviews and fraudulent/legitimate websites. The research explores well-known machine learning classifiers and more effective ensemble models, as well as alternative models, assessed using standard metrics to realize the best-performing ones. Additionally, the dataset is segregated into train, validate, and test representatives with added cross validations to prevent overfitting.

Future research involves improving the accomplishment of the models by refining ensemble models and optimizing hyperparameters to further improve results. This may lead to changes in the ranking of the models based on their performance. Additionally, there is scope for exploring, experimenting with, and improving combinations and comparisons of performance metrics.

Acknowledgement

This paper and the implicated research wouldn't have been possible without the exceptional support of all the co-authors. Sincere thanks to all for their kind assistances and privileged provisions.

Availability of data

The datasets used in the experimentation purposes are available on:
https://www.kaggle.com/datasets/rtatman/deceptive-opinion-spam-corpus and
https://data.mendeley.com/datasets/m7xtkx7g5m/1

References

[1] Allcott, H. and Gentzkow, M., 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), pp.211-236. DOI: https://dx.doi.org/10.1257/jep.31.2.211.

[2] Chesney, R. and Citron, D., 2019. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, *98*, p.147. DOI: https://doi.org/10.1177/14614448241253138.

[3] Shoaib, M.R., Wang, Z., Ahvanooey, M.T. and Zhao, J., 2023, November. Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. In *2023 International Conference on Computer and Applications (ICCA)* (pp. 1-7). IEEE. DOI: https://dx.doi.org/10.1109/ICCA59364.2023.10401723.

[4] Thakur, H.K., Gupta, A., Bhardwaj, A. and Verma, D., 2018. Rumor detection on Twitter using a supervised machine learning framework. *International Journal of Information Retrieval Research (IJIRR)*, *8*(3), pp.1-13. DOI: https://dx.doi.org/10.4018/IJIRR.2018070101.

[5] Chatterjee, R., Pandey, M., Thakur, H.K. and Gupta, A., 2024. Checking Counterfeit Critiques on Commodities using Ensemble Classifiers Enhancing Information Credibility. *Procedia Computer Science*, *233*, pp.570-579. DOI: https://dx.doi.org/10.1016/j.procs.2024.03.246.

[6] Ott, M., Cardie, C. and Hancock, J.T., 2013, June. Negative deceptive opinion spam. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies (pp. 497-501).

[7] Ott, M., Choi, Y., Cardie, C. and Hancock, J.T., 2011. Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557. DOI: https://doi.org/10.48550/arXiv.1107.4557

[8] Islam, M.S., Jyoti, M.N.J., Mia, M.S. and Hussain, M.G., 2023, July. Fake website detection using machine learning algorithms. In 2023 International Conference on Digital Applications, Transformation & Economy (ICDATE) (pp. 255-259). IEEE. DOI: https://dx.doi.org/10.1109/ICDATE58146.2023.10248584.

[9] Jindal, N. and Liu, B., 2007, October. Analyzing and detecting review spam. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 547-552). IEEE. DOI: https://dx.doi.org/10.1109/ICDM.2007.68.

[10] Jindal, N. and Liu, B., 2007, May. Review spam detection. In Proceedings of the 16th international conference on World Wide Web (pp. 1189-1190). DOI: https://dx.doi.org/10.1142/S0219649217500368.

[11] Ott, M., Cardie, C. and Hancock, J., 2012, April. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st international conference on World Wide Web (pp. 201-210). DOI: https://dx.doi.org/10.1145/2187836.2187864.

[12] Li, J., Lv, P., Xiao, W., Yang, L. and Zhang, P., 2021. Exploring groups of opinion spam using sentiment analysis guided by nominated topics. Expert Systems with Applications, 171, p.114585. DOI: https://dx.doi.org/10.1016/j.eswa.2021.114585.

[13] Ren, Y. and Ji, D., 2019. Learning to detect deceptive opinion spam: A survey. IEEE Access, 7, pp.42934-42945. DOI: https://dx.doi.org/10.1007/s10489-022-03427-1.

[14] Adoption rate of emerging technologies in organizations worldwide as of 2020: 2020. https://www.statista.com/statistics/661164/worldwide-cio-surveyoperational-priorities/. Accessed: 2025-05-10.

[15] Weng, H., Li, Z., Ji, S., Chu, C., Lu, H., Du, T. and He, Q., 2018, April. Online e-commerce fraud: a large-scale detection and analysis. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 1435-1440). IEEE. DOI: https://dx.doi.org/10.1109/ICDE.2018.00162.

[16] Jain, A.K. and Gupta, B.B., 2022. A survey of phishing attack techniques, defence mechanisms and open research challenges. Enterprise Information Systems, 16(4), pp.527-565. DOI: https://dx.doi.org/10.1080/17517575.2021.1896786.

[17] Ren, Y., Zhang, L. and Suganthan, P.N., 2016. Ensemble classification and regression-recent developments, applications and future directions. IEEE Computational intelligence magazine, 11(1), pp.41-53. DOI: https://dx.doi.org/10.1109/MCI.2015.2471235.

[18] Shahariar, G.M., Biswas, S., Omar, F., Shah, F.M. and Hassan, S.B., 2019, October. Spam review detection using deep learning. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0027-0033). IEEE. DOI: https://dx.doi.org/10.1109/IEMCON.2019.8936148.

[19] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D., 2019. Text classification algorithms: A survey. Information, 10(4), p.150. DOI: https://dx.doi.org/10.3390/info13020083.

[20] Zhang, M., 2024. Ensemble-Based Text Classification for Spam Detection. Informatica, 48(6). DOI: https://dx.doi.org/10.31449/inf.v48i6.5246.

[21] Asaad, W.H., Allami, R. and Ali, Y.H., 2023. Fake Review Detection Using Machine Learning. Revue d'Intelligence Artificielle, 37(5). DOI: https://dx.doi.org/10.18280/ria.370507.

[22] Sánchez-Paniagua, M., Fidalgo, E., Alegre, E. and Jáñez-Martino, F., 2021. Fraudulent e-commerce websites detection through machine learning. In Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings 16 (pp. 267-279). Springer International Publishing. DOI: https://dx.doi.org/10.1007/978-3-030-86271-8_23.

[23] Kirkpatrick, K., 2016. Battling algorithmic bias: how do we ensure algorithms treat us fairly?. Communications of the ACM, 59(10), pp.16-17. DOI: https://dx.doi.org/10.1145/2983270.

[24] Deceptive Opinion Spam Corpus: 2013. https://www.kaggle.com/datasets/rtatman/deceptive-opinion-spam-corpus. Accessed 2025-05-10.

[25] Fraudulent and Legitimate Online Shops Dataset: 2023. https://data.mendeley.com/datasets/m7xtkx7g5m/1. Accessed 2025-05-10.

[26] What is a Generalized Additive Model?: 2021. https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a. Accessed 2025-05-10.

[27] LASSO (L1) Vs Ridge (L2) Vs Elastic Net Regularization For Classification Model: 2022. https://pub.towardsai.net/lasso-l1-vs-ridge-l2-vs-elastic-net-regularization-for-classification-model-409c3d86f6e9. Accessed 2025-05-10.

[28] Log Loss's Advantage For Determining Overfitting vs AUC: 2025. https://medium.com/@chahnapatel2798/log-losss-advantage-for-determining-overfitting-vs-auc-dad90cb5c61c. Accessed 2025-05-10.

[29] Intuition behind Log-loss score: 2020. https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a/. Accessed 2025-05-10.

[30] Chehal, D., Gupta, P. and Gulati, P., 2023. Predicting the Usefulness of E-Commerce Products' Reviews using Machine Learning Techniques. Informatica, 47(2). DOI: https://dx.doi.org/10.31449/inf.v47i2.4155

[31] How to Develop Elastic Net Regression Models in Python: 2020. https://machinelearningmastery.com/elastic-net-regression-in-python/. Accessed 2025-05-10.

[32] Choosing the Right Metric for Evaluating Machine Learning Models — Part 2: 2018. https://medium.com/usf-msds/choosing-the-right-metric-for-evaluating-machine-learning-models-part-2-86d5649a5428. Accessed 2025-05-10.