International Journal of Information Management Data Insights 3 (2023) 100167

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud

Krishnadas Nanath[a,*], Liting Olney[b]

[a] Middlesex University Dubai, Springs 3, Villa 3, Dubai, United Arab Emirates
[b] Middlesex University Dubai, Knowledge park, Dubai, United Arab Emirates

ARTICLE INFO

Keywords:
Misinformation
Machine learning
Crowdsourcing
Fake content
Online recruitment fraud

ABSTRACT

Misinformation on the web has become a problem of significant impact in an information-driven society. Persistent and large volumes of fake content are being injected, and hence the content (news, articles, jobs, facts) available online is often questionable. This research reviews a range of machine learning algorithms to tackle a specific case of online recruitment fraud (ORF). A model with content features of job posting is tested with five supervised machine learning (ML) algorithms. It then investigates various crowdsourcing techniques that could enhance prediction accuracy and add human insights to machine learning automation. Each crowdsourcing method (explored as human signals online) was tested across the same ML algorithms to test its effectiveness in predicting fake job postings. The testing was conducted by comparing the hybrid models of machine learning and crowdsourced inputs. This study revealed that the best ML algorithm was different in the automated model compared to the hybrid model. Results also indicated that the net promoter type crowdsourced question resulted in the best accuracy in classifying fraudulent and legitimate jobs. The decision tree and generalized linear model demonstrated the highest accuracy among all the tested models.

1. Introduction

With the growth of information available on the internet today, the world faces a severe misinformation problem on the web. Fake content on the internet could come in several forms. It could include spam emails, fake news, fake jobs, fake reviews, and rumors. Social media is no longer just a platform to connect with people; it is heavily used for content creation and sharing. With more than 2 billion monthly active users on Facebook[1] and 300 million on Twitter[2], content sharing becomes very powerful with the reach it generates on these platforms. The absence of control and fact-checking of online content makes social media platforms a fertile ground for misinformation spread (Zubiaga et al., 2018). Therefore, research on the examination of fake content could make smarter systems for creating a safer web.

One category of fake content is the fake jobs that are posted on job portals and professional platforms like LinkedIn.[3] In recent years, online recruitment fraud (ORF) in the form of fraudulent job posts online has increasingly become a serious issue. It has resulted in the misuse of personal information and job applicants' financial loss and harming organizations' credibility (Dutta & Bandyopadhyay;, 2020; Mahbub & Pardede, 2018; Vidros et al., 2017). The Federal Bureau of Investigation (2020) issued a public service announcement in January 2020 stating a considerable increase in ORF since early 2019, with an average loss of $3000 per person. Mahab and Pardede (2018) define online recruitment fraud as "a form of employment scam where a person with fraudulent intentions posts a fake job advertisement on an online platform targeting job seekers." While several studies have explored the use of machine learning algorithms in other forms of fake content detection, research on job postings is scarce.

With the intensive use of analytics in research, several studies use machine learning (ML) algorithms to combat fake content. These algorithms aim to distinguish fake content from non-fake ones that could lead to an automated and adaptive approach (Guzella & Caminhas, 2009). These algorithms learn from an available dataset and train the classification model. They do not rely on human-coded rules that could be subjected to errors and bias. Given a collection of training documents $M_t \in M$ labeled as legitimate or fraudulent, the ML algorithms are learning a function f: M -> {l,f}, for labeling an instance m $\in$ M as legitimate (l) or fraudulent (f). Several studies in the literature compare machine learning algorithms for various topics like bankruptcy

[*] Corresponding author.
E-mail address: username.krishna@gmail.com (K. Nanath).

[1] www.facebook.com

[2] www.twitter.com

[3] www.linkedin.com

https://doi.org/10.1016/j.jjimei.2023.100167
Received 3 January 2022; Received in revised form 8 February 2023; Accepted 13 February 2023

prediction, gesture recognition, and network intrusion (Abdjalil et al., 2010; Barboza et al., 2017; Trigueiros et al., 2012). However, this study attempts to be one of the first to compare the machine learning algorithms for online recruitment fraud (fake job postings).

While machine learning algorithms are useful in fake content detection, they are always vulnerable to adversarial countermeasures by fake content creators (Wang et al., 2014). It is also tough to develop generic models that could be applied to various types of fake content. Therefore, human inputs have been considered influential in enhancing combat with fake content. This led to a fake content detection method via expert verification leading to several third-party fact-checking organizations like Snopes[4] and Factcheck.[5] However, this is a time-consuming activity and involves costs for services trying to discriminate between legitimate and fraudulent content. Therefore, this study explores the idea of crowdsourcing from human signals that they convey on online platforms. These signals could be in the form of reactions, sharing content, recommending articles, or rating the content. This research explored the various types of crowdsourcing signals that could enhance the machine learning approach. A comparison of various machine learning models across the crowdsourcing methods helped develop a hybrid model (human and machine) that was tested for its efficiency in detecting fraudulent jobs. Four approaches of crowdsourcing methods are explored in this research, and the secondary data of job postings is recreated as an online survey for gaining crowdsourcing insights.

The discussion of machine learning for fake content detection, crowdsourcing techniques for human inputs, and the possibility of a hybrid human-machine approach pose several research questions. The complexity of the problem demands novel and reliable solutions. This research is motivated by three crucial research questions (RQs):

RQ1: *Which machine learning algorithms lead to high accuracy and efficiency results in identifying fraudulent job postings online?*

RQ2: *What are the various crowdsourcing options and techniques that could add human inputs to the machine learning model for detecting fake content online?*

RQ3: *In a hybrid human-machine model (machine learning features with crowdsourced inputs), which machine learning algorithms lead to high accuracy and efficiency results in identifying fraudulent job postings online?*

These questions collectively will address the demands of information systems literature to distinguish between fake and real content. Researchers in the past have raised this issue (Michail et al., 2022) in various forms like reviews (Banerjee, 2022), news (Ansar & Goswami, 2021), multimedia content (Kolagati et al., 2022), and others. Studies have highlighted the importance of fake content research in information systems research by not just exploring the methods of detection but also educating internet users. With the growth of content on social media, it is difficult to curtail the propagation of fake content (Thota et al., 2018). Detecting fake content is a challenging area of research; hence, it requires knowledge of various disciplines and novel approaches to advance the field (Shu et al., 2017; Zhou & Zafrani, 2020). A bulk of previous work involves machine learning techniques (Orabi et al., 2020) for fake content detection or developing crowdsourcing platforms that check the veracity of the content (Michail et al., 2022). However, with the growth of technology, there are still various ways for fake content to infiltrate social media (Freitas et al., 2015). Hence, this research uses the best of the two worlds to advance the field of fake content detection using jobs as a case study.

A rigorous research design is used and presented in this study to respond to the above research questions. The rest of this paper is organized as follows: Section 2 reviews related work in the area of fake content detection, crowdsourcing, and hybrid models. Section 3 presents the research method of this study, where three tracks of research design are described. Next, in Section 4, various crowdsourcing techniques are described along with primary data collection details. Section 5 presents the results of machine learning models applied to the secondary and hybrid datasets. Finally, in Section 6, the results are discussed, summarized, and the conclusions are presented.

2. Related work

Misinformation on the web has been an exciting research topic for several researchers in management and technical fields. A literature review conducted by Bondielli and Marcelloni (2019) reveals that the number of papers discussing fake news and rumors has increased almost four times from 2008 to 2018 (Scopus indexed). Based on the topics explored by researchers, they organized misinformation on the web into three categories- fake news, rumors, and others. Fake news refers to articles written to mislead the readers and can be verified as false by other sources (Conroy et al., 2015). The topics explored in fake news research include humorous fakes, social fabrications, and large-scale hoaxes (Rubin et al., 2015). Rumors generally refer to information not confirmed by official sources and spread mostly by social media platforms (DiFonzo & Bordia, 2007; Vosoughi et al., 2017) and eventually leading to virality (Nanath & Joy, 2021). Various categories of rumors are explored in the literature (Knapp, 1944; Zubiaga et al., 2018), but long-standing and breaking news are the most common (Bondielli & Marcelloni, 2019). The literature, however, on fake job postings is scarce.

Machine learning algorithms have been pervasive in detecting fake content (misinformation) online. With various text mining techniques, artificial intelligence algorithms, and human reviews, the literature has shown the effectiveness of various approaches in large-scale fake content detection. The application of machine learning can be seen in three main categories – spam emails (Sharifi et al., 2011), fake news (Katsaros et al., 2019; Nanath et al., 2022), and fake reviews (Hussain et al., 2019). While a few studies used text mining methods like Term Frequency-Inverse Document Frequency (TF-IDF) and Part of Speech tagging (POS), a majority of them used supervised learning approaches like logistic regression (LR), support vector machines (SVM), random forests (RF), decision trees (DT) and naïve Bayes (NB) algorithm. A summary of these papers is provided in Table 1.

While machine learning remains the dominant approach in handling misinformation on the web, very few studies have explored the use of crowdsourcing in combating fake content. The only category of fake content that has explored crowdsourcing as a technique is fake news. Sethi (2017) argues that automated tools cannot verify alternative facts in fake news and propose a crowdsourcing system that leverages human critical thinking to detect fake news with mediation from expert moderators. Pinto et al. (2019) argued that news fact-checking conducted by professional fact-checkers do not scale to the increasing volume of fake news, where crowdsourcing can overcome the limitations and proposed a scalable crowdsourcing process. Although there is a concern about the crowd's wisdom on fact-checking (Hassan et al., 2019), other views are also suggested in the literature. An experiment conducted by Pennycook and Rand (2019) found that human inputs are good at assessing the news source's reliability and suggests that crowdsourcing could be a promising method to improve algorithms in detecting fake news. Other studies have explored crowd signals on Reddit and users' flags on Facebook (Hassan et al., 2019; Tschiatschek et al., 2018)) as inputs for crowdsourcing. These techniques improved the accuracy of fake news detection. Machine learning with crowdsourcing has also been explored in the context of paraphrasing text (Burrows et al., 2013) and predicting data quality (Sheng & Zhang, 2019).

While both machine learning and crowdsourcing have been used in fake content research, there are two prominent literature gaps. First, both these methods have been explored less in the context of online recruitment fraud. Second, these methods have not been used as a hybrid

[4] http://www.snopes.com/

[5] http://factcheck.org/

Table 1
Review of papers using machine learning methods to deal with misinformation on the web.

Category	Paper	Description	Algorithms
Spam email	Sharifi et al. (2011)	Focuses on using LR in detecting internet scams, including spam email, scam queries, and top websites.	LR
Spam email	Hassan and Mtetwa (2018)	Discussion on features supervised ML classifiers and performance metrics on datasets for detecting spam emails	NB, SVM
Spam email	Dada et al. (2019)	Review of ML algorithms for classifying emails into spam and not spam.	Density based clustering, K-nearest neighbor (KNN), NB, Nur, Firefly, Rough set, SVM, DT, NBTree, C4.5/J48, Ensemble classifiers, RF
Spam email	Suryawanshi et al. (2019)	Focuses on spam email detection and classification using ML classifiers, along with performance evaluation.	Naive Bayes, SVM, KNN, Adaboost, and Ensemble Classifiers
Spam email	Nandhini and Marseline (2020)	Performance comparison on ML classification algorithm for spam email detection.	LR, DT, NB, KNN, and SVM
Fake news	Katsaros et al. (2019)	Performance evaluation of eight ML algorithms for fake news detection and classification.	LR, C-Support Vector, Gaussian naive Bayes, Multinomial naive Bayes, DT, RF, Multilayer perceptron, Convolutional neural networks
Fake news	Asr and Taboada (2019)	Discussion on the text classification problem in detecting fake news, including NLP and feature-based models.	
Fake news	Reis et al. (2019)	Focuses on feature extraction for fake news detection and performance evaluation of classifiers.	KNN, NB, RF, SVM, XGBoost
Fake news	Ozbay and Alatas (2020)	Proposes a two-step method in detecting fake news on online social media, focusing on data pre-processing and 23 supervised algorithms.	BayesNet, JRip, OneR, Decision Stump, ZeroR, Stochastic Gradient Descent, CV Parameter Selection, Randomizable Filtered Classifier, Logistic Model Tree, Locally Weighted Learning, and others.
Fake news	Zhang and Ghorbani (2020)	Literature review of fake news detection to date, covering research-based approaches including ML models and feature selection.	DT, RF, SVM, LR, KNN
Fake review	Banerjee et al. (2015)	Focuses on using Supervised Learning to Classify Authentic and Fake Online Reviews, using ten machine learning algorithms for analysis.	LR, C4.5, back-propagation network, JRip, NB, RF, SVML, SVMP, SVMRBF, and Voting.
Fake review	Hassan and Islam (2019)	Focusing on semi-supervised and supervised learning for detecting fake online reviews using a dataset.	NB, SVM
Fake review	Hussain et al. (2019)	Literature review of existing 76 studies on spam review detection, focusing on pre-processing, feature extraction, and machine learning approaches.	SVM, NB, DT, LR, Rule-based
Fake review	Martens and Maalej (2019)	Focuses on detecting fake reviews on the Apple App Store experimented on a balanced and imbalanced dataset.	RF, DT, Multilayer perceptron, Linear support vector classification, Gaussian NB

approach to combating fake content. There are few studies in the recruitment fraud domain, and most of them use a publically available Employment Scam Aegean Dataset (EMSCAD), created by Vidros et al. (2017). Mahbub and Pardede (2018) proposed the inclusion of contextual features in job postings in addition to textual and structural information used by Vidros et al. (2017) to improve model performance. The behavioral features were added by Nindyati and Nugraha (2019) to improve the model performance in recruitment fraud detection. The algorithms in ORF included ensemble techniques (Lal et al., 2019), random forests (Alghamdi & Alharby, 2019), and Naïve Bayes (Mahbub & Pardede, 2018). This research aims to address the gap by using a mix of machine learning and crowdsourcing approaches (various combinations) to detect online recruitment fraud and suggest the best approach.

3. Research method

The research method of this study has been designed considering the approaches of various research papers that have explored big data analytics and machine learning. The motive of big data analytics is to use the volume, variety, and veracity of data to arrive at actionable insights (Kushwaha et al., 2021). While the volume of the data might not be huge in this paper, there are complexities involved in the variety and veracity of data. Machine learning algorithms are explored in one part of this study to learn from past data to improve measurable performance in these tasks (Ray, 2019). It is essential to experiment with ML algorithms as standard implementation makes it challenging to extract meaningful information and gather knowledge (Kar, 2016; Chakraborty & Kar, 2017). Research method in this study is divided into – data collection, an overview of three tracks, and data analysis.

3.1. Data collection

Given the objectives of this research, a secondary dataset would be helpful to start the analysis. A good dataset of already coded fraudulent and non-fraudulent jobs could set up the research design for comparing machine learning approaches in online recruitment fraud. Hox & Boeije, 2005 guidelines were followed for using secondary research, and the challenges were evaluated. The Employment Scam Aegean Dataset (EMSCAD) (LICS, 2017) was used for this study as it contains real-life job ads posted by Workable.[6] The dataset has been used in the past (Alghamdi & Alharby, 2019; Vidros et al., 2017) and has proven to be a practical testbed in exploring the space of ORF.

The dataset originally contains 17,880 job postings, out of which 17,014 are legitimate while other postings are fraudulent. The dataset is reliable as all the postings in this dataset were manually annotated by specialized employees from Workable. Several quality checks, like suspicious activity from the client, false contracts, wrong company information, and user complaints, ensured the correct annotation of fraud jobs in the dataset. The records in the dataset were a mix of structured and unstructured data. The fields in the dataset were of four types- string, HTML fragment, binary and nominal. A summary of the fields is provided in Table 2. A record in the dataset could be described as:

Set of fields F= {F1,F2,….Fn} $_{n=16,}$

and a binary class field C {+,-} $_{\text{job posting is fraudulent or not}}$

[6] http://workable.com/

Table 2
Summary of fields in the EMSCAD dataset.

Number	Variable	Description	Type
1	job_id	Unique Job ID	identifier
2	title	The title of job ad entry	string
3	location	Geographical location of the job ad	string
4	department	Corporate department (e.g. sales).	string
5	salary_range	Indicative salary range (e.g. $50,000–$60,000)	string
6	company_profile	A brief company description.	text (html fragment)
7	description	The details description of the job ad.	text (html fragment)
8	requirements	Enlisted requirements for the job opening.	text (html fragment)
9	benefits	Enlisted offered benefits by the employer.	text (html fragment)
10	telecommuting	True for telecommuting positions.	binary
11	has_company_logo	True if company logo is present.	binary
12	has_questions	True if screening questions are present.	binary
13	employment_type	Full-type, Part-time, Contract, etc.	norminal (6 levels)
14	required_experience	Executive, Entry level, Intern, etc	norminal (8 levels)
15	required_education	Doctorate, Master's Degree, Bachelor, etc.	norminal (14 levels)
16	industry	Automotive, IT, Health care, Real estate, etc.	norminal (132 levels)
17	function	Consulting, Engineering, Research, Sales etc.	norminal (38 levels)
18	fraudulent	target - Classification attribute.	binary

One of the problems associated with the dataset is class imbalance (95% legitimate and 5% fraud). Lal et al. (2019) describe this as a major problem in ORF detection, and several researchers have approached this problem by creating a more balanced dataset (Dutta & Bandyopadhyay, 2020; Nindyati & Nugraha, 2019; Vidros et al., 2017). A new subset of the original dataset was created with stratified sampling to maintain a 70–30% ratio of legitimate and fraudulent job postings. An overall subset of 1000 postings was created. Besides solving the imbalance problem, this subset would allow crowdsourcing inputs to be added using survey data.

3.2. Research method overview- three tracks

The research method of this study is divided into three tracks. The first track is titled 'secondary data analysis,' and the initial step of this trick is data preparation (as described in Section 3.1) and coming up with a balanced subset of 1000 job postings. In finalizing the dataset, several steps were taken to clean the data treatment of missing values, examine inappropriate data points, and handle invalid attributes. The second step involved selecting existing variables (Table 2) and creating new variables (Table 5) for the analysis that could help predict fraudulent job postings. The insights for this step came from exploratory data analysis that helped gain insights on feature selection. This track's last step involved executing and comparing machine learning algorithms (described in Section 3.3) on the reduced dataset.

The second track is titled 'identifying crowdsourcing techniques.' Crowdsourcing attempts to add human inputs into the model to see if the fraudulent jobs detection accuracy could be enhanced. There are various methods of asking respondents about the quality of job postings. One of the methods is directly posting the question- Is this job posting a fraudulent one? However, such questions cannot be used in job portals and social media as it would be an added responsibility on users, and many would choose not to respond. Therefore, this study explores various crowdsourcing methods in the literature and uses the same dataset to add more variables via crowdsourcing techniques (Section 4). The data was collected using a questionnaire, and the primary data collection is described in Section 4.4.

The third track is titled 'hybrid approach.' The secondary dataset was combined with the crowdsourced data to form a hybrid dataset. The hybrid dataset is then subjected to various machine learning algorithms to observe any accuracy changes and investigate the best crowdsourcing method from the given options. Various studies in the literature have suggested that a hybrid approach could enhance the evaluation parameters of the machine learning models (Nielek et al., 2016; Shabani & Sokhn, 2018). The results of machine learning algorithms on various crowdsourcing methods are provided in Section 5.3. A summary of the three tracks as a research method overview is provided in Fig. 1.

The focus of research methods is unique in each track provided in Fig. 1. While the first track (secondary data) focuses on data preparation and ML models for establishing the based model, the second track (crowdsourcing techniques) proposes various methods of getting crowdsourcing inputs. These inputs are based on theoretical foundations from the literature and extend the previous work on combining human inputs with ML models. One output of the first track feeds into the evaluation stage, while the second output is fed into the second track as the fake jobs are converted into questionnaires that could gather human input. The research method focus in the third track is on implementing ML models in the hybrid model (crowdsourcing and base model). The first two tracks provide input to this track, while its output is fed into the evaluation stage. Each track is discussed further in detail.

3.3. Data analysis

Machine learning algorithms in this research have been used to address two research questions- RQ1 and RQ3. The first research question compares various machine learning algorithms using only the secondary dataset. In contrast, the third research question compares them across various crowdsourcing techniques on the hybrid dataset (secondary data and crowdsourced primary data). This study compares five machine learning algorithms for binary classification for the detection of fraudulent and non-fraudulent job postings. These algorithms were chosen based on their popularity in fraud detection algorithms and their applicability in the context of this research. The five algorithms were logistic regression, decision tree, random forests, naïve Bayes, and generalized linear model.

Logistic regression is a regression model designed to test the logit (log-odds) probability of the dependent variable – fraudulent vs. legitimate job postings. It has been used widely in the context of fraud detection (Patil et al., 2018). The relationship between the dependent variable and the independent variables used as features of the job posting is measured using the logistic function defined as,

$$g(z) = 1/(1 + e^{-z}),$$

The regression equation based on the feature vector X = (Feature$_1$, Feature$_2$, ...,Feature$_n$) of the job postings and the dependent variable y (fraudulent = 1, non-fraudulent = 0) can be represented as:

$$Logit(y) = \beta_0 + \sum_{i=1}^{n} \beta_i Feature_i$$

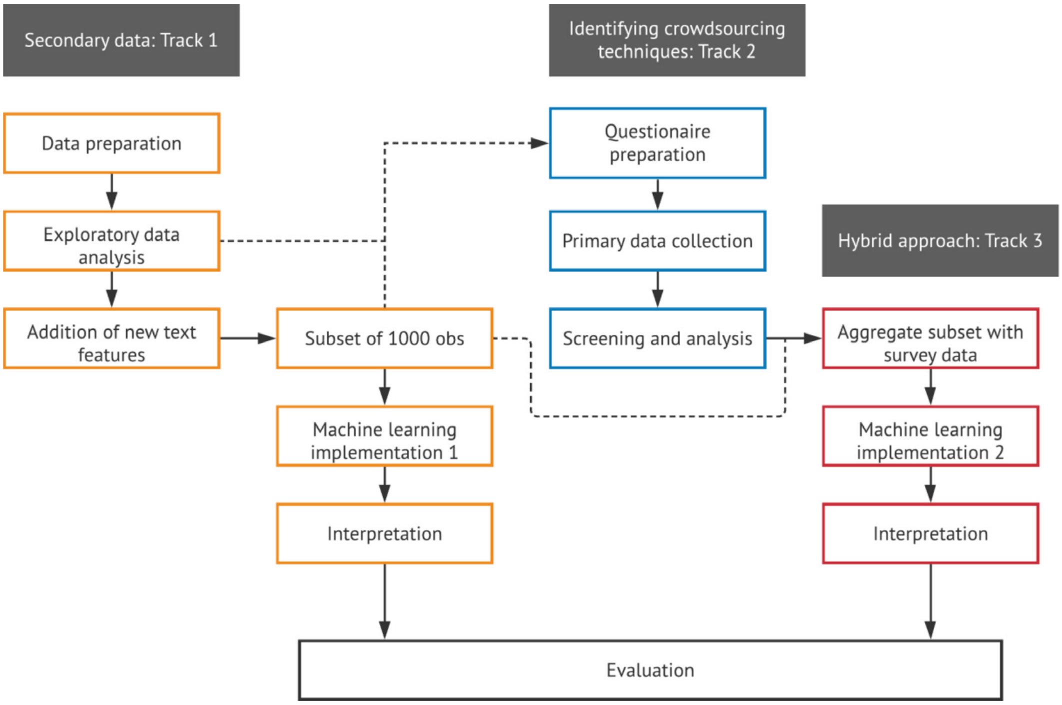

Secondary data: Track 1
Identifying crowdsourcing techniques: Track 2
Hybrid approach: Track 3
Data preparation
Exploratory data analysis
Addition of new text features
Subset of 1000 obs
Machine learning implementation 1
Interpretation
Questionaire preparation
Primary data collection
Screening and analysis
Aggregate subset with survey data
Machine learning implementation 2
Interpretation
Evaluation

Fig. 1. Research method overview of three tracks.

Naïve Bayes algorithm is based on the Bayes theorem and allows probabilistic understanding to be learned. It has shown high predictive performance for a classification task (John & Langley, 2013), and its statistical components can be formally expressed (Han and Kamber, 2006). The simplest form of the Bayesian network classifier is a simple Bayes classifier (Duda & Hart, 2006). Given a training set D ={x_i, t_i} (i = 1 to n) with input vectors xi= {x_{i1}, x_{i2}, ….x_{in}} and target labels $t_i \in \{0,1\}$ (fraudulent=1, non-fraudulent=0), it follows that,

$$p(x|t) = \prod_{m=1}^{n} p(x^{(m)}\ |t)$$

On the other hand, the decision tree algorithm is an algorithm that requires fewer assumptions about the data distribution. It is a non-parametric classification algorithm, and it classifies a sample input by passing it through the tree and allocating it to the appropriate leaf node. A Gini index is usually used in the CART algorithm to implement a decision tree, and it measures the impurity of the dataset to find the splitting criterion. If D is the training set that contains m class labels C_i, the Gini index (Han et al., 2012) can be represented as,

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

where p_i denotes the probability that a tuple in D belongs to class C_i. Random forest is considered an extension of the decision tree algorithm that combines the random subspace feature selection and bagging to merge single decision trees (Breiman, 2001). The generalized linear model has also been an important algorithm in classification problems and is also used in fraud recognition. It was developed by Nelder and Wedderburn (1972) and is considered more flexible than simple regression. The relationship between the dependent and independent variable is constructed by a link function such as log or power. GLM is known to have a more comprehensive application range than simple regression and can obtain a relationship model that is closer to reality (Kao et al., 2011).

4. Crowdsourcing techniques and primary data collection

Crowdsourcing is a useful method that has been explored for detecting fake content. It falls under the knowledge-based category of fake content detection described by Shu et al. (2017). There are two approaches to human fact-checking suggested by Shu et al. (2017)- export-oriented and crowdsourcing-oriented. Expert-oriented fact-checking relies on human domain experts to verify the facts online (Bondielli & Marcelloni, 2019; Stahl, 2018). However, this method is time-consuming, intellectually very demanding, and could involve high costs when executed on large datasets. On the other hand, the crowdsourcing method explores the crowd's wisdom to gain more insights into the content and subject it to fake detection tests.

The crowdsourcing method has been widely used in the literature of fake news detection (Gravanis et al., 2019). Pennycook and Rand (2019) found that human inputs are useful sources of discerning news reliability. It was also suggested that crowdsourcing could increase fake news detection efficiency (Tschiatschek et al., 2018). Since pure crowdsourcing could raise the issue of reliability and costs (Shabani & Sokhn, 2018), it could be interesting to observe the hybrid approach's results (machine learning and crowdsourcing). This section attempts to review various crowdsourcing methods (CSMs) that exist in the literature (to address RQ2) and select the techniques that can be used for this research.

4.1. *The direct approach to fake content detection (CSM 1)*

A common practice in crowdsourcing is to directly ask the respondents to check if the content is fake. This could involve asking questions to the respondents or users of an online platform regarding the legitimacy of the information's source and accuracy. The response could either be binary or categorical. This method has been used in a few studies (Pennycook & Rand, 2019; Ghadiyaram & Bovik, 2015) to fight misinformation on social media. A summary of these studies and the nature of crowdsourcing questions are provided in Table 3.

While this method is useful for research and creating a dataset, it might not be a practical question to ask in platforms that display

Table 3
Summary of direct approach crowdsourcing.

Crowdsourcing question format	Question	Studies adopting this format
Categorical	Do you think that the above job post is from a legitimate source? (Yes / I don't know / No)	Pennycook and Rand (2019), Ghadiyaram and Bovik (2016), and Costa et al. (2011).
Binary	The above content contains accurate information. (True/False)	Tchakounte et al. (2020) and Hung et al. (2017).

content like news, jobs, or articles. Such questions could be an added responsibility for newsreaders, social media users, or job seekers. It could prevent the service providers from using these questions on their portal. Further, studies have demonstrated in the literature that direct questions could result in potential bias (Roni et al., 2020; Stockemer, 2019). Therefore, several indirect methods could be explored for seeking insights from the respondents, and these methods could be used in online platforms. These methods could be part of the online platform, and the responses can be used for testing the hybrid approach.

4.2. The net promoter score approach (CSM 2)

Net promoter score (NPS) is a technique that was introduced in the marketing space to measure customer satisfaction (Reichheld, 2003) and has been widely used by several organizations. The question used to capture this score is usually applied in the context of a product, service, or organization. The nature of the question appears in the format of "*How likely is it that you would recommend [Product/Service/Company X] to a friend or colleague?*" The question is answered on a scale of zero to ten, where ten indicates "extremely likely" to recommend, five means neutral, and zero means "not at all likely.".

This type of crowdsourcing approach has been termed as a social signal by Vedova et al. (2018). They classify the approaches to fake news detection as content-based methods and social signals. While the former method looks at the content of the fake news by deploying techniques like natural language processing, the latter explores human signals to gain additional insights. The signals could also include engagement and interaction of users on posts that are put up on social media. One such signal is how much respondents would be interested in recommending content (news, job, or an article) to their peers. This approach follows the question format of NPS, and the modified version of this method has been used in various studies (Alhabash & McAlister, 2015; Catallo & Martinenghi, 2017; Vedova et al., 2018).

While several organizations and research papers have adopted NPS, it has also faced criticism from several researchers (Brandt, 2007). They have questioned the theoretical foundation of NPS and the categorization of the original rating scale. While there are concerns over its use in the marketing context, it would be interesting to test this in the hybrid approach using machine learning. In the context of fake job postings, the question format is modified as *"How likely is it that you would recommend this job posting to a friend, colleague or your network?"*.

4.3. The fuzzy logic approach (CSM 3)

The third type of crowdsourcing approach attempts to deviate from using scales to measure the signals. For several years, fuzzy logic has been used to embed expert inputs into computer models for a broad range of applications (Aburrous et al., 2010). It is based on the principle that the majority of things in the world are uncertain and are characterized by two traits –fuzzy and random (Zadeh, 1965). It suggests that there is a possibility of something being a member of a set through the membership function, and the values could range between 0 and 1. If x is a member of X, and A is a fuzzy set with the membership function being μ_A, the fuzzy set A can be written as:

$$A = \{(x, \mu_A(x)),\ x\ \in X\},\ \mu_A(x) :\ X\ \to [0,1]$$

It has proven to be a useful alternative for measuring risks and fraud detection (Shah, 2003). It provides more information to effectively assess the users' response to the crowdsourcing activity than the qualitative and scale-based responses. Studies in the past have shown that fuzzy logic could improve the Likert scale to measure the variables. Therefore, the response format can be changed to a number format between 0 and 1 (instead of an n-point scale). A few attempts have been seen in the literature to use this crowdsourcing format to attempt fake content detection (Shabani & Sokhn, 2018; Song et al., 2018). The questions have either explored the reliability of the posting or the legitimacy of the content. Since this research avoids direct fact verification, reliability was chosen. The question format in the context of the job posting would appear as "*How would you rate the reliability of this job posting?*" The response would be in a slider format that varies from 0 to 1, with values closer to 0 indicating low reliability, while those closer to 1 indicating high reliability.

4.4. Engagement-based approach (CSM 4)

Engagement on social media platforms has been an exciting source of study to take in human reactions to online content. Several platforms are extending their range of engagements to seek a variety of inputs. Facebook extended its list of reactions from Like and Love to more options like 'Haha,' 'Wow,' 'Sad,' and 'Angry.' Several online portals use star ratings (1 to 5 stars) to assess user satisfaction and reaction on a product, service, or online content. The range of engagement options has been considered an effective way of data collection (Tian et al., 2017). Interestingly, the 'Angry' and 'Sad' reactions have been used to explore news articles' misinformation on social media (Masullo & Kim, 2021).

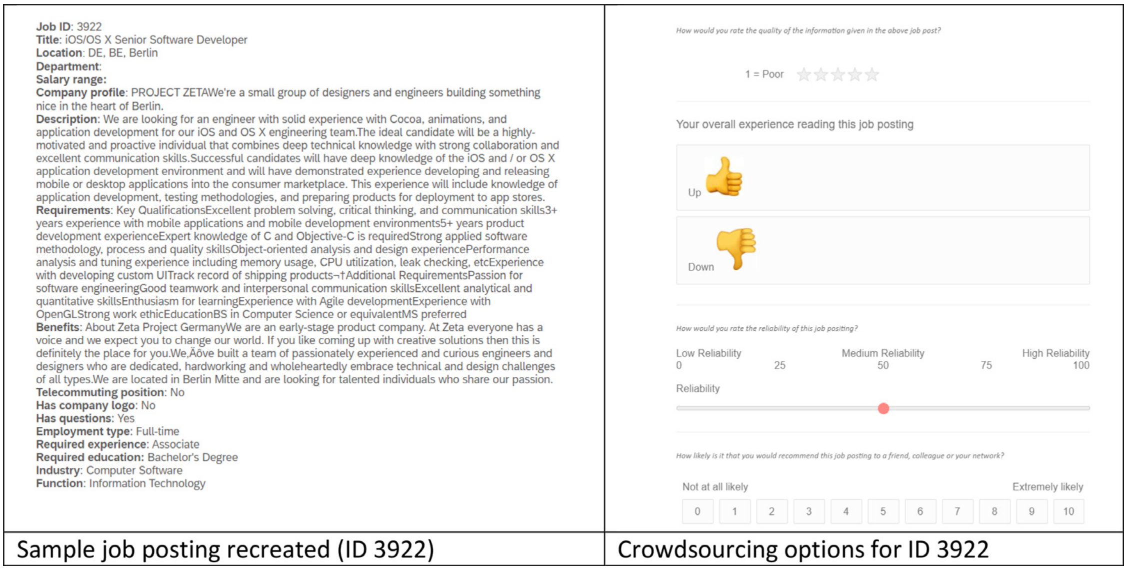
Since this research applies to social media and online portals, a generic approach for engagement was adopted as a crowdsourcing strategy. Two approaches were tested as part of the user engagement-ratings (star-based) and like/dislike (binary). The star ratings were inspired by the Likert scale used in several studies that have explored fake content detection via surveys (Ghadiyaram & Bovik, 2015; Chatterjee et al.,2017; Welinder & Perona, 2010). The question format for this method would appear as "*How would you rate the quality of the information given in the above job post?*" This question's response would vary from 1 star (poor) to 5 stars (excellent). The second approach of binary response is inspired by studies that have used binary responses like True and False to gauge information on the online content's reliability (Hung et al., 2017; Tchakounte et al., 2020). The question format for this method would appear as, "Your overall experience reading this job posting." The response would be in the form of Thumbs up (1) or Thumbs down (0). A summary of all the methods used in this research for hybrid modeling is provided in Table 4.

4.5. Primary data collection – crowdsourcing design

A survey method was chosen to gather the crowdsourced intelligence on the fraud job posting dataset. Each job in the reduced EMSCAD dataset (1000 job postings) was recreated on an online survey tool (Qualtrics). There were three parts to the survey. The first part provided instructions to the respondents to read the job postings and answer the questions. The research's objective was not communicated as it could result in potential bias and suspicion when answering the

Table 4
Summary of the crowdsourcing methods.

Method	Question	Response	Related studies
The net promoter score approach *(CSM 2)*	*How likely is it that you would recommend this job posting to a friend, colleague or your network?*	Scale [0 to 10]: 10- extremely likely to recommend, 5- neutral, and 0- not at all likely.	Alhabash and McAlister (2015), Catallo and Martinenghi (2017), and Vedova et al. (2018)
The fuzzy logic approach *(CSM 3)*	*How would you rate the reliability of this job positing?*	Slider format, probability range from 0 to 1.	Song et al. (2018), Shabani and Sokhn (2018).
Engagement based approach *(CSM 4)*	[a] *How would you rate the quality of the information given in the above job post?*	Star rating [1 to 5]: 1-poor, and 5- excellent.	Ghadiyaram and Bovik (2016), Welinder and Perona (2010), Chatterjee et al. (2017), Sethi (2017), and Simpson et al. (2015)
	[b] Your overall experience reading this job posting	Binary (0 or 1): 1- Thumbs up, and 0- Thumbs down	Tchakounte et al. (2020), and Hung et al. (2017).



Job ID: 3922
Title: iOS/OS X Senior Software Developer
Location: DE, BE, Berlin
Department:
Salary range:
Company profile: PROJECT ZETAWe're a small group of designers and engineers building something nice in the heart of Berlin.
Description: We are looking for an engineer with solid experience with Cocoa, animations, and application development for our iOS and OS X engineering team.The ideal candidate will be a highly-motivated and proactive individual that combines deep technical knowledge with strong collaboration and excellent communication skills.Successful candidates will have deep knowledge of the iOS and / or OS X application development environment and will have demonstrated experience developing and releasing mobile or desktop applications into the consumer marketplace. This experience will include knowledge of application development, testing methodologies, and preparing products for deployment to app stores.
Requirements: Key QualificationsExcellent problem solving, critical thinking, and communication skills3+ years experience with mobile applications and mobile development environments5+ years product development experienceExpert knowledge of C and Objective-C is requiredStrong applied software methodology, process and quality skillsObject-oriented analysis and design experiencePerformance analysis and tuning experience including memory usage, CPU utilization, leak checking, etcExperience with developing custom UITrack record of shipping products¬†Additional RequirementsPassion for software engineeringGood teamwork and interpersonal communication skillsExcellent analytical and quantitative skillsEnthusiasm for learningExperience with Agile developmentExperience with OpenGLStrong work ethicEducationBS in Computer Science or equivalentMS preferred
Benefits: About Zeta Project GermanyWe are an early-stage product company. At Zeta everyone has a voice and we expect you to change our world. If you like coming up with creative solutions then this is definitely the place for you.We,Äôve built a team of passionately experienced and curious engineers and designers who are dedicated, hardworking and wholeheartedly embrace technical and design challenges of all types.We are located in Berlin Mitte and are looking for talented individuals who share our passion.
Telecommuting position: No
Has company logo: No
Has questions: Yes
Employment type: Full-time
Required experience: Associate
Required education: Bachelor's Degree
Industry: Computer Software
Function: Information Technology

Sample job posting recreated (ID 3922)

How would you rate the quality of the information given in the above job post?

1 = Poor ☆☆☆☆☆

Your overall experience reading this job posting

Up 👍

Down 👎

How would you rate the reliability of this job positing?

Low Reliability 0 | 25 | Medium Reliability 50 | 75 | High Reliability 100

Reliability

How likely is it that you would recommend this job posting to a friend, colleague or your network?

Not at all likely — Extremely likely

0 1 2 3 4 5 6 7 8 9 10

Crowdsourcing options for ID 3922

Fig. 2. Sample job posting and crowdsourcing options.

questions. The second part of the survey listed ten job postings that were recreated from the EMSCAD subset data. The job posts contained all associated attributes listed in the dataset, except for the class label variable, to provide the participants with as much information as available. The third part of the survey reflected the crowdsourcing exercise where the four crowdsourcing methods (CSM 2, CSM 3, CSM 4.a, and CSM 4.b) provided in Table 4 were listed for each job posting in the survey. A sample job posting, along with the crowdsourcing questions, is shown in Fig. 2.

One hundred surveys were created and administered using Qualtrics to cover all the thousand job postings in the dataset. Since each survey listed ten job postings and each posting was clubbed with four crowdsourcing questions, a total of 40 questions were posted in each survey. The idea of restricting one survey with only ten job postings was to avoid burnout and fatigue components of the respondents that could affect the responses' quality (Minnaar & Heystek, 2013). A pilot study was conducted with five participants to pretest the questionnaire to check if the respondents understood the questions as intended and to test the response format before the primary data collection (Blair et al., 2013; Saris & Gallhofer, 2014). Since the subset maintained a 70–30 ratio of legitimate and fraudulent job postings, the same ratio was maintained in each survey. Therefore, every survey of 10 job postings had seven legitimate job postings and three fraudulent postings. The postings were randomized by the survey tool. The respondents were chosen based on convenience sampling; however, similar demographic variables (age, education, experience) were maintained to avoid reporting bias. A total of 100 respondents responded to the surveys, covering the crowdsourcing exercise of 1000 job postings.

5. Results: comparison of machine learning approaches across various models

This section presents the results of machine learning models to compare the algorithms (RQ1) and to compare the effectiveness of various crowdsourcing techniques (RQ3). Feature selection techniques allowed the authors to set a base model for comparing the ML algorithms. The base model represents the model without any inputs from the crowdsourcing methods. This model could provide a baseline to check whether human inputs could improve the results of prediction. Also, various ML algorithms can be tested on the base model to understand the performance within the base model and then between other models. Four techniques used for crowdsourcing were added to the base model (one at a time) as an additional independent variable resulting in four models. These models were then compared across various machine-learning algorithms.

5.1. Feature selection

A large number of experiments were conducted to select the best features for developing a predictive model using RapidMiner and visualization packages in R. The top predictors of the model were selected using insights from the tools and inputs from experts. The same experts who validated the survey design helped provide inputs of variables that could be important for fraudulent job detection as they were domain experts.

Crosstab visualizations were meaningful in selecting important independent variables for the model. A sample visualization in Fig. 3 indicates that the majority (66%) of job posts in the fraudulent class

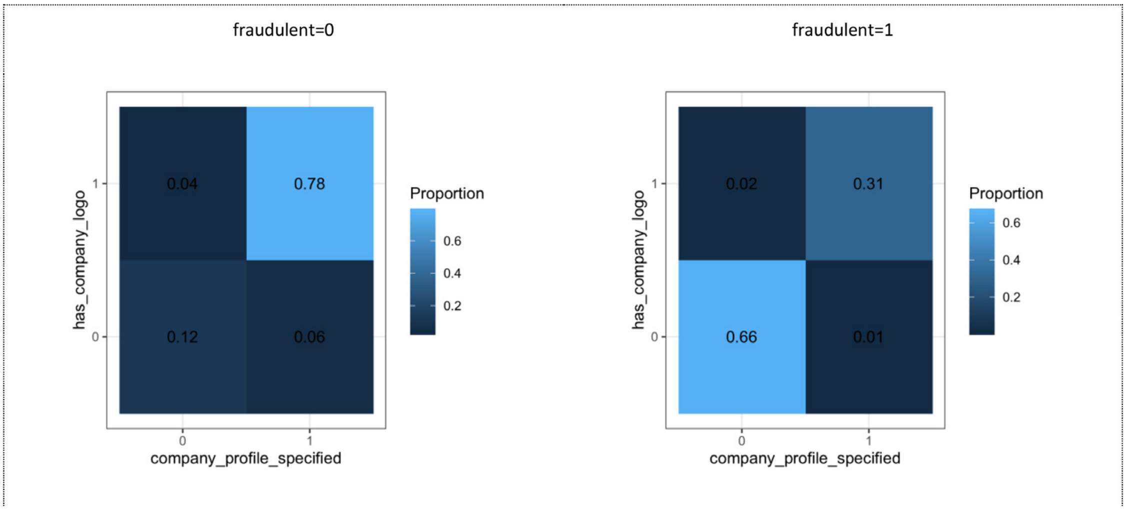


Fig. 3. Sample crosstab visualization for fraudulent and legitimate jobs across the company log and company profile.

Table 5
A summary of variables used in the predictive model.

Model variables	Description	Type	Existing or new feature	Summary Statistics
company_profile_specified	TRUE if company profile is specified in the job posting, FALSE otherwise.	Binary	New feature	TRUE (699), FALSE (301)
requirements_specified	TRUE if requirements is specified in the job posting, FALSE otherwise.	Binary	New feature	TRUE (835), FALSE (165)
benefits_specified	TRUE if benefits are stated in the job description, FALSE otherwise.	Binary	New Feature	TRUE (586), FALSE (414)
salary_specified	TRUE if the salary is specified in the job description, FALSE otherwise.	Binary	New Feature	TRUE (167), FALSE (833)
location_specified	TRUE if the location is specified in the job description, FALSE otherwise.	Binary	New Feature	TRUE (977), FALSE (23)
company_profile_word	Word count in the company profile section of the job posting. Larger values indicate more detailed information about the company.	Numeric	New feature	Min: 0, Max: 424, Mean: 80.3
description_word	Word count in the description section of the job posting. Larger values indicate a more detailed job description.	Numeric	New feature	Min: 1, Max: 1184, Mean: 171.6
requirements_word	Word count in the requirements section of the job posting. Larger values indicate a more detailed demand for the requirements.	Numeric	New feature	Min: 0, Max: 623, Mean: 74.23
benefits_word	Word count in the benefits section of the job posting. Larger values indicate more details on the job benefits.	Numeric	New feature	Min: 0, Max: 362, Mean: 29.38
has_company_logo	TRUE if company logo is present in the job posting, FALSE otherwise	Binary	Existing variable	TRUE (690), FALSE (310)
has_questions	TRUE if screening questions are present in the job posting, FALSE otherwise.	Binary	Existing variable	TRUE (443), FALSE (557)
emptype_specified	TRUE if the employment type is specified in the job posting (Full-time, Part-time, and others), FALSE otherwise.	Binary	New feature	TRUE (805), FALSE (195)
reqedu_specified	TRUE if the required education is specified in the job posting (Doctorate, Bachelors, and others), FALSE otherwise.	Binary	New feature	TRUE (568), FALSE (432)

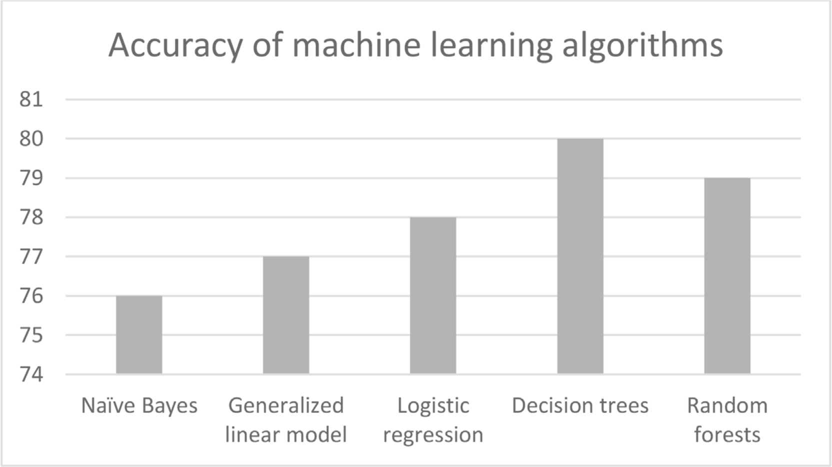
have no company logo or company profile. The majority (78%) of legitimate jobs have both features, on the other hand. Therefore, the presence of the company logo and the company profile (binary variables- Yes or No) served as essential features for detecting a fraudulent job posting. Similar visualization techniques and feature information on correlation (with target variable), identical values within the feature, and stability from RapidMiner helped develop the final list of features for the predictive model. While some variables existed in the original dataset, most of the features were newly derived from the existing data. A summary of the variables used for machine learning modeling is provided in Table 5.

5.2. Results of machine learning algorithms for EMSCAD subset

Five algorithms were implemented on the EMSCAD subset (1000 job postings)- logistic regression, decision tree, random forests, naïve Bayes, and generalized linear model. The crowdsourcing inputs were not introduced at this stage so that the change in parameters can be observed once they are added to the predictive model. The algorithms were implemented in the same tool (RapidMiner) so that not only the accuracy and model parameters can be observed, but other parameters like execution time can be compared as well.

The overall accuracy of the model was calculated with the help of the confusion matrix. There are four components of this matrix- number of correctly identified fraudulent job postings, the number of correctly identified legitimate job postings, the number of jobs identified as fraudulent (but were not), and the number of articles not identified as fraudulent (while they were). The algorithms' computational time was also captured to gain insights into the scaled-up implementation of the project. The highest accuracy was observed for the decision tree algorithm (80%), while naïve Bayes took the shortest computational time (and scoring time for 1000 rows). All five algorithms demonstrated a reasonably consistent accuracy range (76– 80%), with the lowest accuracy obtained from naïve Bayes. The computational time of random forests was the highest, but it was expected as it involves constructing a multitude of decision trees when training the model. A comparison of prediction accuracy is provided in Fig. 4, while the computational and scoring time is summarized in Fig. 5.

Besides accuracy, several evaluation parameters are derived from the confusion matrix. They serve as important parameters in assessing the strength of the algorithms. Four additional parameters were considered for performance evaluation – recall, precision, F measure, and AUC. Recall measure represents the capture rate and can be calculated



,

Fig. 4. Accuracy of five machine learning algorithms in the job posting subset (without crowdsourcing).

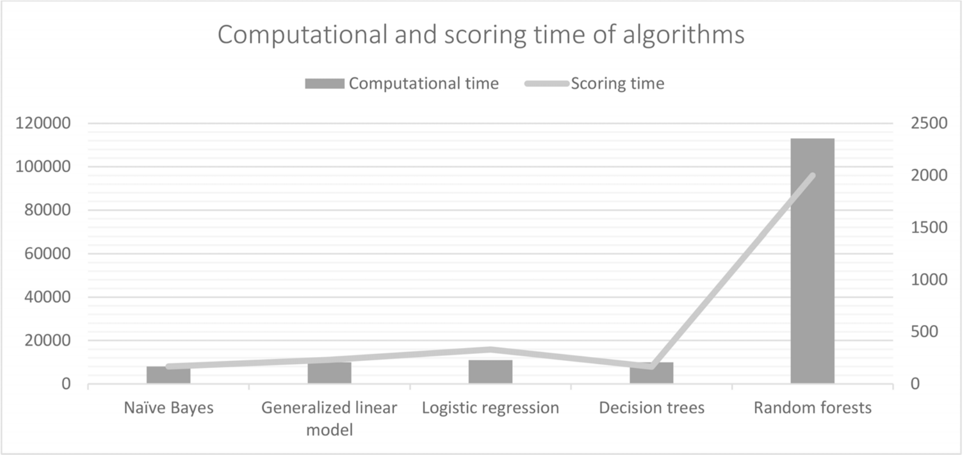


Fig. 5. Computational and scoring times of five machine learning algorithms in the job posting subset (without crowdsourcing).

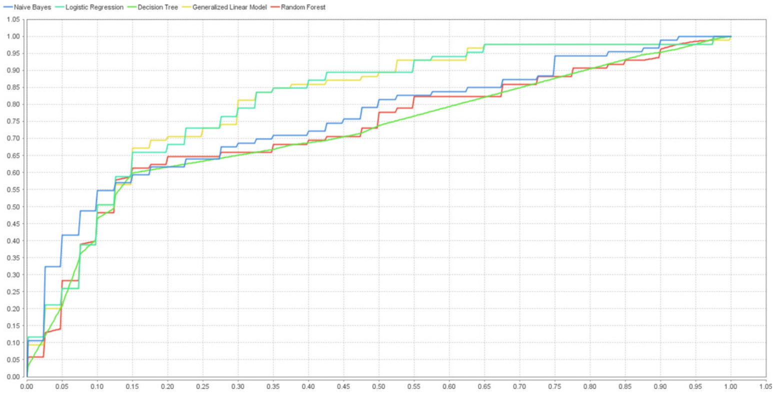


Fig. 6. ROC comparison plots of five machine learning algorithms.

by examining the ratio of correctly predicted positive examples by total positive examples in the dataset. On the other hand, precision represents the hit rate and is defined as the ratio of correctly predicted positive cases to the number of cases labeled by the model as positive. F measure combines recall and precision to examine the relationship between the positive labels and those given by the classifier.

The last parameter of the performance evaluation (AUC) is calculated using ROC plots. These plots compare the true positive rate versus the false positive rate for various classification threshold values. ROC comparison plot for various machine learning algorithms is presented in Fig. 6. The area under the curve is called the AUC, and the best discriminating model will have an area index of 1 (Cui et al., 2008). A summary of the parameters used for the evaluation is presented in Table 6.

Table 6
Summary of evaluation parameters.

Parameter	Method
Precision	True Positive / (True Positive +False Positive)
Recall	True Positive / (True Positive + False Negative)
F Measure	2* (Precision*Recall)/ (Precision+ Recall)
AUC	The area under the ROC curve

While the highest accuracy was demonstrated by the decision tree algorithm (80%), the highest AUC value was obtained for logistic regression and the generalized linear model (0.81). The hit rate was the highest for the decision tree algorithm. However, a combined perspective (with recall and F measure) suggests logistic regression as an

Table 7
Results of evaluation parameters for five machine learning models.

	Naïve Bayes		Logistic regression		Decision tree		Random forest		Generalized linear model	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
AUC	0.75	*0.02*	0.81	*0.02*	0.72	*0.03*	0.72	*0.03*	0.81	*0.02*
Precision	0.60	*0.05*	0.63	*0.05*	0.70	*0.05*	0.68	*0.03*	0.63	*0.04*
Recall	0.59	*0.04*	0.60	*0.03*	0.56	*0.10*	0.55	*0.09*	0.55	*0.05*
F Measure	0.60	*0.02*	0.61	*0.03*	0.62	*0.07*	0.61	*0.05*	0.59	*0.05*

* SD=Standard Deviation.

Table 8
Summary of crowdsourcing models for comparison.

Model	Description	Number of features	Crowdsourcing Method
Base Model	Predictive model with 13 features given in Table 5	13	n/a
Model 1	Predictive model with 13 features of the base model and CSM 2 variable (net promoter method, scale of 0 to 10).	14	CSM 2
Model 2	Predictive model with 13 features of the base model and CSM 3 variable (fuzzy logic approach, numeric range from 0 to 1).	14	CSM 3
Model 3	Predictive model with 13 features of the base model and CSM 4.a variable (engagement based star rating, Likert scale 1 to 5).	14	CSM 4.a
Model 4	Predictive model with 13 features of the base model and CSM 4.b variable (engagement based like/dislike, binary variable with values 1 and 0).	14	CSM 4.b

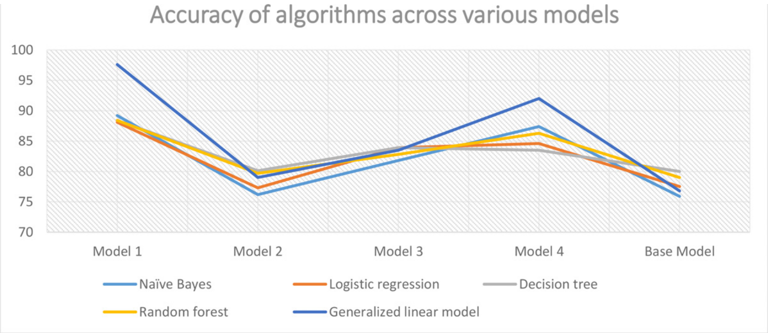


Fig. 7. Prediction accuracy of ML algorithms across five models.

effective method for classifying fraudulent and legitimate job postings. The results of the evaluation parameters for the five machine-learning methods are given in Table 7.

5.3. Results of the crowdsourcing models

The model evaluated in the previous section served as a base to compare various crowdsourcing techniques. Each crowdsourcing technique provided in Table 4 (CSM 2, CSM 3, CSM 4.a, and CSM 4.b) served as one variable that could be added to the predictive model. These variables were treated as one additional variable, each adding to the list of thirteen independent variables given in Table 5. Therefore, four new models were run with fourteen independent variables (thirteen from the base model and one crowdsourced). These models were then compared on accuracy and the four performance evaluation parameters given in Table 6. A summary of the models is given in Table 8.

All five machine learning algorithms were implemented across four models. Compared to the base model, most crowdsourcing methods demonstrated an increase in prediction accuracy (except Model 2). The highest accuracy in general across all methods was observed in Model 1 (CSM 2, net promoter score) followed by Model 4 (CSM 4.b, engagement-based approach, overall experience binary). A common factor between the two models was the overall experience and recommendation factor. The variables of these models also demonstrated significant correlation as people who rated an NPS score of 6 and above were very likely to give a thumbs up (value 1) as the overall experience. Model 2 (CSM 3, reliability fuzzy logic approach) demonstrated the lowest accuracy among the crowdsourcing techniques. Therefore, the net promoter score approach and engagement type approach turned out to be better crowdsourcing methods to improve prediction accuracy. When the machine learning algorithms were compared, the decision tree and generalized linear model demonstrated the highest accuracy among all the tested models. A comparison graph of accuracy across models using five machine learning algorithms is provided in Fig. 7. The evaluation parameters of the same exercise are given in Table 9.

6. Discussion

The first round of comparisons without the crowdsourcing insights (base model) revealed exciting insights. The objective of base model implementation was to compare the five machine learning algorithms in predicting fraudulent job listings and serve as a baseline model to compare the crowdsourcing methods. It is a known fact that accuracy is not the only parameter measuring the effectiveness of a machine learning algorithm, and it was demonstrated by the base model results (Fig. 4 and Table 7). While the decision tree turned out to have the best accuracy, it did not demonstrate good AUC results. In terms of computational time, naïve Bayes took the least time for execution and scoring but was compromised in classification accuracy.

The comparison of machine learning algorithms for a secondary dataset has been explored by various studies (Liu et al., 2017; Osisanwo et al., 2017; Zhang et al., 2017;). The evaluation parameters are also used. However, the only consistent result compared to this research is the almost inverse relationship between accuracy and computational time. The experiment conducted by Zhang et al. (2017) revealed that the top-performing algorithm (in terms of accuracy), like the ran-

Table 9
Comparison of hybrid models across five machine learning algorithms.

		Naïve Bayes	Logistic regression	Decision tree	Random forest	Generalized linear model
Model 1 (CSM 1)	AUC	0.97	0.99	0.99	0.99	0.99
	Precision	0.78	0.72	0.73	0.73	0.93
	Recall	0.89	0.99	0.99	0.99	0.99
	F Measure	0.83	0.84	0.84	0.84	0.96
Model 2 (CSM 3)	AUC	0.76	0.82	0.80	0.86	0.84
	Precision	0.61	0.65	0.72	0.87	0.68
	Recall	0.63	0.52	0.58	0.38	0.59
	F Measure	0.61	0.58	0.63	0.53	0.62
Model 3 (CSM 4.a)	AUC	0.84	0.88	0.80	0.88	0.88
	Precision	0.69	0.77	0.71	0.73	0.72
	Recall	0.73	0.66	0.74	0.72	0.77
	F Measure	0.70	0.71	0.72	0.72	0.74
Model 4 (CSM 4.b)	AUC	0.94	0.88	0.86	0.86	0.98
	Precision	0.95	0.75	0.72	0.83	0.96
	Recall	0.62	0.73	0.75	0.69	0.78
	F Measure	0.75	0.74	0.73	0.75	0.86

dom forest, had very slow training time efficiency and, hence, slower execution time. However, all other results comparing accuracy, AUC, and execution time were very different from previous studies using the same machine learning algorithms. The second part of the research explored crowdsourcing as a technique for detecting fraudulent job postings. Four crowdsourcing models were explored in this study. A comparison of four models (presented in Table 9) revealed that various formats of questions influence prediction accuracy. Overall, the hybrid approach (secondary features and primary crowdsourced data) revealed better accuracy than the base model (secondary only). This could be an exciting insight for future researchers building predictive models for fake content detection.

Comparing the four crowdsourcing methods, Model 1 (CSM 2, net promoter score) demonstrated the highest accuracy. There was a jump of sixteen percent in predictive accuracy with the top-performing machine learning algorithm. The two top-performing crowdsourcing methods were the net promoter score (scale of 0 to 10) and engagement-based approach (overall experience, binary variable). A common feature in both these methods was the focus on experience and recommendation of the content. These methods did not probe the legitimacy of the demonstrated content. There was a significant correlation between the methods, as people who are likely to recommend the job to their network would also have a good experience reading the content. However, the range of inputs (0 to 10) allowed capturing more variations when compared to a binary response (1 or 0), and hence higher accuracy was observed.

The lowest-performing method across all five ML algorithms was Model 2 (CSM 3, fuzzy logic approach). There were only marginal increments of accuracy compared to the base model in this method. There could be two reasons for lower accuracy- the nature of the response scale (probability between 0 and 1) and the nature of the question asked for the crowdsourcing exercise. In the context of fake job postings, the latter appears to be a valid reason, as the nature of the question probed the reliability of the job content. This closely mirrored the first crowdsourcing method (CSM 1, direct approach), where respondents are directly asked to flag the fake content. While a direct approach could be time-consuming, and users might not respond on a live platform like social media or a job portal, it could also introduce bias in answering the question. Hence, this research adds a unique contribution to the body of literature investigating fraud detection using crowdsourcing methods.

6.1. Theoretical implications

This research is one of the first research to classify various crowdsourcing methods using an online platform for improving machine learning models. Several studies have explored secondary data to investigate fake content online, and the crowdsourcing inputs use user reviews (Harris, 2012), flagging activity by users (Tschiatschek et al., 2018), e-commerce reviews (Kaghazgaran et al., 2017), tweets (Ansar & Goswami, 2021) and others. However, limited studies are exploring the type of questions that could be used for crowdsourcing insights. This study explored the effect of various questions (that users could respond to) on the power of discriminating fraudulent versus legitimate job postings. This classification can be used by future researchers to compare and improve their machine-learning models with human inputs. These methods can easily be incorporated into online questionnaires and social media platforms.

Another theoretical implication is the support of this study in developing a theoretical framework for the comparison of machine learning algorithms. In previous studies, the comparison has been conducted in various domains – bankruptcy prediction (Barboza et al., 2017), malicious webpages detection (Kazemian & Ahmed, 2015), gesture recognition (Trigueiros et al., 2012), network intrusion (Abdjalil et al., 2010), disk failure prediction (Pitakrat et al., 2013) and insurance sector (Rawat et al., 2021). However, this research would be one of the firsts in exploring machine learning comparisons for fake job postings. Dyson and Golab (2017) explore the comparisons in fake news detection, and this study enters the limited space of research available in comparing ML algorithms in detecting fake online content. Based on the review of these studies, it is clear that there is no clear winner when it comes to prediction accuracy, and hence such research should be explored for particular topics rather than generic datasets.

6.2. Implications for practice

While the study has theoretical implications in comparing machine learning models for fraud detection, it has practical implementations for content providers. Social media platforms like Facebook have been continuously using several methods to stop the spread of fake content and false propaganda (Brown, 2018). Facebook has explored several methods to combat fake content on its platform, and the only crowdsourcing method explored so far is direct reporting (Facebook, 2020). This research could be the starting point to explore techniques beyond direct reporting and using other crowdsourcing signals to gain insights into fraudulent content. This research has concrete implications for job portals and professional platforms like LinkedIn. These online portals can introduce more options (beyond the apply button) to understand user experience and use them as crowdsourcing signals to detect fraudulent jobs. Overall, the theoretical and practical implications combined can lead to a framework that could be used for future research in fake job predictions. This framework is presented in Fig. 8.

6.3. Limitations

As with other studies, there are several limitations to this research. First, default parameters of the machine learning algorithms were used

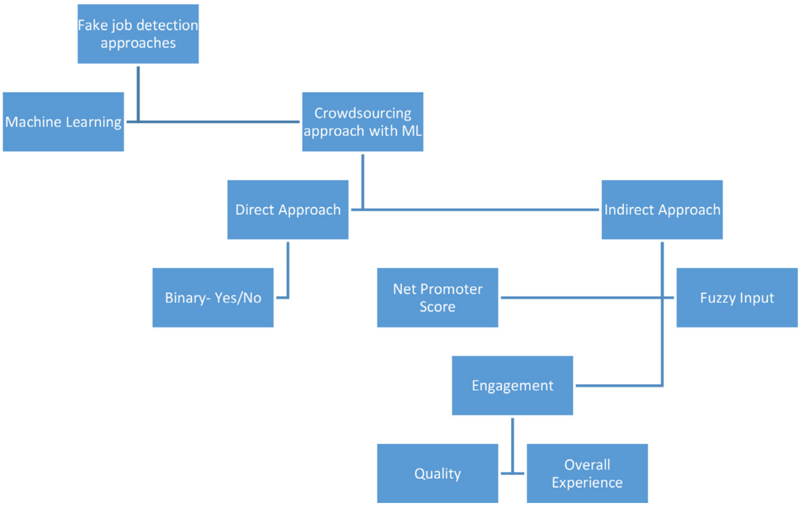


Fig. 8. A framework for future research on fake job prediction.

as they were implemented in RapidMiner. Hence, the full capacity of the models could not be utilized. Second, there was not enough depth in the feature selection. However, this is common in studies exploring fraud detection and fake content prediction (Barboza et al., 2017). Since feature selection in such problems could involve individual judgment, the theoretical basis becomes less reliable (Pal et al., 2016). Third, the crowdsourcing exercise could not be executed at a larger scale as it involves time and resources beyond this study's capacity. More extensive crowdsourced data could help gain more insights into big data analytics.

References

Abdjalil, K., Kamarudin, M. H., & Masrek, M. N. (2010). Comparison of machine learning algorithms performance in detecting network intrusion. In *Proceedings of the international conference on networking and information technology* (pp. 221–226). IEEE.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert Systems with Applications, 37*(12), 7913–7921.

Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security, 10*(3), 155–176.

Alhabash, S., & McAlister, A. R. (2015). Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter. *New media & society, 17*(8), 1317–1339.

Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights, 1*(2), Article 100052.

Torabi Asr, F., & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society, 6*(1), Article 2053951719843310.

Banerjee, S. (2022). Exaggeration in fake vs. authentic online reviews for luxury and budget hotels. *International journal of information management, 62*, Article 102416.

Banerjee, S., Chua, A. Y., & Kim, J. J. (2015). Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th international conference on ubiquitous information management and communication* (pp. 1–7).

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417.

Blair, J., Czaja, R. F., & Blair, E. A. (2013). *Designing surveys: A guide to decisions and procedures*. Sage Publications.

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences, 497*, 38–55.

Brandt, D. R. (2007). For good measure-on the one number you need to grow, one size doesn't fit all. *Marketing Management, 16*(1), 20.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32.

Brown, J. (2018, March 29). Facebook's plan to fight election interference includes weeding out fake memes and videos. Gizmodo. Retrieved February 7, 2023, from https://gizmodo.com/facebooks-plan-to-fight-election-interference-includes-1824189286.

Burrows, S., Potthast, M., & Stein, B. (2013). Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 4*(3), 1–21.

Catallo, I., & Martinenghi, D. (2017). The dimensions of crowdsourcing task design. In *Web engineering: Proceedings of the 17th international conference, ICWE 2017, Rome, Italy, June 5–8, 2017,* (pp. 394–402). Springer International Publishing.

Chakraborty, A., & Kar, A. K. (2017). Swarm intelligence: A review of algorithms. *Nature-Inspired Computing and Optimization: Theory and Applications*, 475–494.

Chatterjee, S., Mukhopadhyay, A., & Bhattacharyya, M. (2017). Judgment analysis based on crowdsourced opinions. In *Proceedings of the IEEE 33rd international conference on data engineering (ICDE)* (pp. 1439–1443). IEEE.

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4.

Costa, J., Silva, C., Antunes, M., & Ribeiro, B. (2011). On using crowdsourcing and active learning to improve classification performance. In *Proceedings of the 11th international conference on intelligent systems design and applications* (pp. 469–474). IEEE.

Cui, G., Leung Wong, M., Zhang, G., & Li, L. (2008). Model selection for direct marketing: Performance criteria and validation methods. *Marketing Intelligence & Planning, 26*(3), 275–292.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon, 5*(6), e01802.

DiFonzo, N., & Bordia, P. (2007). Rumor, gossip and urban legends. *Diogenes, 54*(1), 19–35.

Duda, R. O., & Hart, P. E. (2006). *Pattern classification*. John Wiley & Sons.

Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology, 68*(4), 48–53.

Dyson, L., & Golab, A. (2017). Fake news detection exploring the application of nlp methods to machine identification of misleading news sources. CAPP 30255 Adv. Mach. Learn. Public Policy.

Facebook. (2020). Working to stop misinformation and false news. Working to Stop Misinformation and False News | Meta for Media. Retrieved February 7, 2023, from https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. (2015). Reverse engineering socialbot infiltration strategies in twitter. In *Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining 2015* (pp. 25–32).

Ghadiyaram, D., & Bovik, A. C. (2015). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing, 25*(1), 372–387.

Gravanis, G., Vakali, A., Diamantaras, K., & Karadais, P. (2019). Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications, 128*, 201–213.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications, 36*(7), 10206–10222.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques third edition*. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

Han, J., & Kamber, M. (2006). Classification and prediction. *Data mining: Concepts and techniques*, 347–350.

Harris, C. G. (2012). Detecting deceptive opinion spam using human computation. In *Proceedings of the workshops at the twenty-sixth AAAI conference on artificial intelligence* In.

Hassan, M. A., & Mtetwa, N. (2018). Feature extraction and classification of spam emails. In *Proceedings of the 5th international conference on soft computing & machine intelligence (ISCMI)* (pp. 93–98). IEEE.

Hassan, N., Yousuf, M., Mahfuzul Haque, M. A., Suarez Rivas, J., & Khadimul Islam, M. (2019). Examining the roles of automation, crowds and professionals towards sustainable fact-checking. In *Proceedings of the companion world wide web conference* (pp. 1001–1006).

Hassan, R., & Islam, M. R. (2019). Detection of fake online reviews using semi-supervised and supervised learning. In *Proceedings of the international conference on electrical, computer and communication engineering (ECCE)* (pp. 1–5). IEEE.

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary versus secondary. In K Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 593–599). Atlanta, GA: Elsevier Science.

Hung, N. Q. V., Thang, D. C., Tam, N. T., Weidlich, M., Aberer, K., Yin, H., et al. (2017). Answer validation for generic crowdsourcing tasks with minimal efforts. *The VLDB Journal, 26*, 855–880.

Hussain, N., Turab Mirza, H., Rasool, G., Hussain, I., & Kaleem, M. (2019). Spam review detection techniques: A systematic literature review. *Applied Sciences, 9*(5), 987.

John, G.H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. arXiv preprint arXiv:1302.4964.

Kaghazgaran, P., Caverlee, J., & Alfifi, M. (2017). Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond. In *Proceedings of the international AAAI conference on web and social media: 11* (pp. 560–563). Vol.No..

Kao, W. K., Chen, H. M., & Chou, J. S. (2011). Aseismic ability estimation of school building using predictive data mining models. *Expert Systems with Applications, 38*(8), 10252–10263.

Kar, A. K. (2016). Bio inspired computing–a review of algorithms and scope of applications. *Expert Systems with Applications, 59*, 20–32.

Katsaros, D., Stavropoulos, G., & Papakostas, D. (2019). Which machine learning paradigm for fake news detection? In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence* (pp. 383–387).

Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications, 42*(3), 1166–1177.

Knapp, R. H. (1944). A psychology of rumor. *Public Opinion Quarterly, 8*(1), 22–37.

Kolagati, S., Priyadharshini, T., & Rajam, V. M. A. (2022). Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. *International Journal of Information Management Data Insights, 2*(1), Article 100054.

Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights, 1*(2), Article 100017.

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble learning based online recruitment fraud detection. In *Proceedings of the twelfth international conference on contemporary computing (IC3)* (pp. 1–5). IEEE.

LICS, Laboratory of Information and Communication Systems, University of the Aegean, Samos, Greece. EMSCAD Employment Scam Aegean Dataset, 2017. Available online: Http://icsdweb.aegean.gr/emscad (accessed on 22 February 2017).

Liu, Y., Bi, J. W., & Fan, Z. P. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications, 80*, 323–339.

Mahbub, S., & Pardede, E. (2018). Using contextual features for online recruitment fraud detection.

Martens, D., & Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering, 24*(6), 3316–3355.

Masullo, G. M., & Kim, J. (2021). Exploring “angry” and “like” reactions on uncivil Facebook comments that correct misinformation in the news. *Digital Journalism, 9*(8), 1103–1122.

Michail, D., Kanakaris, N., & Varlamis, I. (2022). Detection of fake news campaigns using graph convolutional networks. *International Journal of Information Management Data Insights, 2*(2), Article 100104.

Minnaar, L., & Heystek, J. (2013). Online surveys as data collection instruments in education research: A feasible option? *South African Journal of Higher Education, 27*(1), 162–183.

Nanath, K., & Joy, G. (2021). Leveraging Twitter data to analyze the virality of COVID-19 tweets: A text mining approach. *Behaviour & Information Technology, 42*(2), 1–19.

Nanath, K., Kaitheri, S., Malik, S., & Mustafa, S. (2022). Examination of fake news from a viral perspective: An interplay of emotions, resonance, and sentiments. *Journal of Systems and Information Technology, 24*(2), 131–155.

Nandhini, S., & KS, J. M. (2020). Performance evaluation of machine learning algorithms for email spam detection. In *Proceedings of the international conference on emerging trends in information technology and engineering (ic-ETITE)* (pp. 1–4). IEEE.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General), 135*(3), 370–384.

Nielek, R., Georgiew, F., & Wierzbicki, A. (2016). Crowd teaches the machine: reducing cost of crowd-based training of machine classifiers. In *Proceedings of the part II 15 artificial intelligence and soft computing: 15th international conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016* (pp. 502–511). Springer International Publishing.

Nindyati, O., & Nugraha, I. G. B. B. (2019). Detecting scam in online job vacancy using behavioral features extraction. In *Proceedings of the international conference on ICT for smart society (ICISS)* (pp. 1–4). IEEE. Vol. 7.

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management, 57*(4), Article 102250.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128–138.

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications, 540*, Article 123174.

Pal, R., Kupka, K., Aneja, A. P., & Militky, J. (2016). Business health characterization: A hybrid regression and support vector machine analysis. *Expert Systems with Applications, 49*, 48–59.

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia Computer Science, 132*, 385–395.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, 116*(7), 2521–2526.

Pinto, M. R., de Lima, Y. O., Barbosa, C. E., & de Souza, J. M. (2019). Towards fact-checking through crowdsourcing. In *Proceedings of the IEEE 23rd international conference on computer supported cooperative work in design (CSCWD)* (pp. 494–499). IEEE.

Pitakrat, T., Van Hoorn, A., & Grunske, L. (2013). A comparison of machine learning algorithms for proactive hard disk drive failure detection. In *Proceedings of the 4th international ACM sigsoft symposium on architecting critical systems* (pp. 1–10).

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), Article 100012.

Ray, S. (2019). A quick review of machine learning algorithms. In *Proceedings of the international conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35–39). IEEE.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–55.

Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems, 34*(2), 76–81.

Roni, S. M., Merga, M. K., & Morris, J. E. (2020). *Conducting quantitative research in education*. Berlin/Heidelberg, Germany: Springer.

Rubin, V. L., Chen, Y., & Conroy, N. K. (2015). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM conference on hypertext and social media* (pp. 315–316).

Shabani, S., & Sokhn, M. (2018). Hybrid machine-crowd approach for fake news detection. In *Proceedings of the IEEE 4th international conference on collaboration and internet computing (CIC)* (pp. 299–306). IEEE.

Shah, S. (2003). *Measuring operational risks using fuzzy logic modeling*. Towers Perrin-Tillinghast.

Sharifi, M., Fink, E., & Carbonell, J. G. (2011). Detection of internet scam using logistic regression. In *Proceedings of the IEEE international conference on systems, man, and cybernetics* (pp. 2168–2172). IEEE.

Sheng, V. S., & Zhang, J. (2019). Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence: 33* (pp. 9837–9843). Vol.No..

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

Simpson, E. D., Venanzi, M., Reece, S., Kohli, P., Guiver, J., Roberts, S. J., et al., (2015). Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proceedings of the 24th international conference on world wide web* (pp. 992–1002).

Song, J., Wang, H., Gao, Y., & An, B. (2018). Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems, 159*, 244–258.

Stahl, K. (2018). Fake news detection in social media. *California State University Stanislaus, 6*, 4–15.

Stockemer, D. (2019). *Quantitative methods for the social sciences a practical introduction with examples in spss and stata. by the registered company*. Springer Nature Switzerland AG.

Suryawanshi, S., Goswami, A., & Patil, P. (2019). Email spam detection: An empirical comparative study of different ml and ensemble classifiers. In *Proceedings of the IEEE 9th international conference on advanced computing (IACC)* (pp. 69–74). IEEE.

Tchakounté, F., Faissal, A., Atemkeng, M., & Ntyam, A. (2020). A reliable weighting scheme for the aggregation of crowd intelligence to detect fake news. *Information, 11*(6), 319.

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. *SMU Data Science Review, 1*(3), 10.

Tian, Y., Galery, T., Dulcinati, G., Molimpakis, E., & Sun, C. (2017). Facebook sentiment: Reactions and emojis. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 11–16).

Trigueiros, P., Ribeiro, F., & Reis, L. P. (2012). A comparison of machine learning algorithms applied to hand gesture recognition. In *Proceedings of the7th Iberian conference on information systems and technologies (CISTI 2012)* (pp. 1–6). IEEE.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Proceedings of the companion proceedings of the the web conference 2018* (pp. 517–524).

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., & De Alfaro, L (2018). Automatic online fake news detection combining content and social signals. In *Proceedings of the 22nd conference of open innovations association (FRUCT)* (pp. 272–279). IEEE.

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(1), 6.

Vosoughi, S., Mohsenvand, M. N., & Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD), 11*(4), 1–36.

Wang, G., Wang, T., Zheng, H., & Zhao, B. Y. (2014). Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *Proceedings of the 23rd {USENIX} security symposium ({USENIX} security 14)* (pp. 239–254).

Welinder, P., & Perona, P. (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition-workshops* (pp. 25–32). IEEE.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications, 82*, 128–150.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management, 57*(2), Article 102025.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR), 53*(5), 1–40.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR), 51*(2), 1–36.