An Intelligent Arabic Model for Recruitment Fraud Detection Using Machine Learning

Mohamed A. Sofy[1]*, Mohammed H. Khafagy[2], and Rasha M. Badry[1]

[1] Information System Department, Faculty of Computers and Information, Fayoum University, Fayoum 63514, Egypt; Email: rmb01@fayoum.edu.eg

[2] Computer Science Department, Faculty of computers and Information, Fayoum University, Egypt; Email: mhk00@fayoum.edu.eg

*Correspondence: ma3152@fayoum.edu.eg

Abstract—Over the last years, with the tremendous growth of digital transformation and the constant need for companies to hire employees, huge amounts of fraudulent jobs have been posted on the internet. A cleverly planned sort of scam aimed at job searchers for a variety of unprofessional purposes is a false job posting. It can lead to a loss of money and effort. An Arabic intelligent model has been built to avoid fraudulent jobs on the Internet using machine learning, data mining, and classification techniques. The proposed model is applied to the Arabic version of the EMSCAD dataset. It is available on the Internet in the English version and it has been retrieved from the use of a real-life system and consists of several features such as company profile, company logo, interview questions, and more features depending on job offer ads. Firstly, EMSCAD is translated into the Arabic language. Then, a set of different classifiers such as Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbor (KNN) was used to detect the fraudulent jobs. Finally, the results were compared to determine the best classifier used for detecting fraudulent jobs. The proposed model achieved better results when using a Random Forest classifier with 97% accuracy.

Keywords—data mining, fraud detection, online recruitment, machine learning, EMSCAD dataset

I. Introduction

The trend towards technology requires us to constantly research the development of our services. Companies have relied on working through the Internet (working from home). The recruitment processes have become very easy to do online through many recruitment portals such as "Wazef", "Forasna", where job seekers upload their resumes and skills. Companies check their files and contact the candidates to schedule an online interview. There have been 174 reported cases of employment scams across Australia during the month of February 2022, alone. Among 174 reported cases, 16.1% incurred financial loss and the total amount adds up to 142,762 Australian Dollars [1].

This huge number makes ambitions tend towards data theft, taking advantage of the youth's need for work. Using computers as an instrument to further illegal ends, such as committing fraud, intellectual property, stealing identities, or violating privacy. Cybercrime, especially through the Internet, has grown in importance as the computer has become central to commerce and entertainment.

Cybercrimes constitute many problems for the individual and society, as the cost of damage around the world is estimated at 6 trillion dollars by the year 2021 [2]. Therefore, we need to prevent the theft of confidential data to avoid theft. To solve this issue, prevention, and detection method can be used.

Egypt's Vision 2030 is heading towards digital transformation and automation in various fields of work, raising the degree of Flexibility and competitiveness of the economy, increasing employment rates, and decent job opportunities, and improving the business environment, and society. Therefore, it was necessary to follow up and detect fraudulent recruitment operations, the purpose of which is to steal individuals' data and threaten them, as criminals publish fake advertisements for employment and exploit job seekers. This causes the loss of money and data for individuals, as well as the loss of the reputation of the organization and the threat to society.

Machine learning is one of the most important solutions and has many examples in our daily lives, such as Netflix and Siri, Companies also use machine learning to create perceptions about future visions, improve customer service and reduce costs, but most of this data is available in languages other than Arabic, which may be English or French. It was necessary to reduce these crimes in our Arab region, and therefore we need a huge volume of data in Arabic and the lack of data. It was important to translate the available dataset into the Arabic language, this is considered a new challenge in our research paper, as it does not exist before. The difficulty of working and understanding the Arabic language is because it contains many synonyms, in addition to its many types which are traditional Arabic, as in the language of the Qur'an, the modern classical language, used in official conversations and on television, and the vernacular. Therefore, the ambiguity in the language makes the Arabic language difficult to learn for machines to make decisions.

In addition, it has become very beneficial to use data mining and data analysis to predict and detect fraudulent

Manuscript received July 17, 2022; revised September 13, 2022; accepted September 28, 2022; published February 17, 2023

jobs. So, this paper aims to develop an intelligent model form to detect fraudulent jobs on the Internet using machine learning algorithms. Four machine learning algorithms were used: Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM).

Random forest is a learning algorithm that is supervised. The "forest" is a collection of decision trees that are typically trained using the "bagging" method to improve the overall output. RF creates many decision trees and blends them to generate a more precise and reliable prediction. As a result, it can be applied to both classification and regression issues [2, 3, 4].

A decision tree or a bagging classifier has approximately the same hyper parameters as a random forest. There is no need to combine a decision tree with a bagging classifier because, after using this approach, you may just utilize the random forest classifier class, we discovered a significant improvement in results [3].

The decision tree Algorithm is a supervised machine learning algorithm. It can be used for both classification problems as well as for regression problems. We developed a model that predicts the value of a target variable (fraudulent), in which the decision tree solves the problem using the tree representation, with the leaf node corresponding to a class label (fraudulent) and attributes represented on the internal node of the tree. Any of the branches of that decision node can identify the results of those tests. Starting from the beginning at the root and this tree are going through until a leaf node is reached. It is the way of obtaining classification results from a decision tree [3].

Because it is simple to develop and particularly helpful for very large datasets, NB is a classification strategy based on the Bayes Theorem with the assumption of independence between predictors. This classification follows its method, which consists of several steps [5]:

- Convert the data set into a frequency table
- Creates a table of probabilities by finding the probabilities
- Use the Naive Bayesian equation to calculate the post-probability of each category. The layer with the highest posterior probability is the prediction result.

K-Nearest Neighbor is a Machine Learning algorithm based on the Supervised Learning technique. It assumes that the new data and existing cases are similar. Then, the available categories place the new example in the category that is the most comparable to the others. This means that as fresh data comes, the KNN algorithm can quickly classify it into a well-suited category [6].

Support Vector Machine is yet another popular modern machine learning approach [7]. Support-vector machines in machine learning are supervised learning models with related learning algorithms that examine data used for regression and classification analyses.

SVMs can effectively do non-linear classification in addition to linear classification by implicitly mapping their inputs into high-dimensional feature spaces. This technique is known as the kernel trick. It essentially draws lines between the classes. The margins are drawn to have the shortest possible distance between them and the closest data points, which minimizes classification errors [8].

We used The Employment Scam Aegean (EMSCAD) dataset [9], which constitutes a huge number of real and fake ads. We need to use different features and unconventional and unused methods to reach the intelligent model and improve results.

The remainder of this paper is structured as follows: Section II, presents the related works on fraud job detection; Section III presents the proposed model. Thereafter, we give details about the methodology. Sections IV and V discuss the features weight and Classification techniques. In Section VI, we present the results of machine learning classification algorithms, and Section VII presents a paper discussion. Finally, we conclude and outline some possible future works in Section VIII.

A classifier maps input variables to target classes using training data. The classifiers used in the paper to differentiate fake job ads from others are briefly discussed. There are two forms of classifier-based predictions: single classifier predictions and ensemble classifier predictions.

II. Related Work

Previous studies relied on (EMSCAD) dataset. It contains real-life job ads posted by Workable, which includes nearly 17,000 advertisements in English. No research work was applied to the Arabic fraud job, while the research was directed towards Arabic fake news. In this research [10], the authors were collecting data and news from YouTube through comments. This research applied three machine learning algorithms such as Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB). SVM achieved an accuracy of 95.35%. While the DT achieved 93.47%. And finally, the NB achieved 92.38%. The best accuracy is by SVM, when all the rumor topics are mixed, the best results in terms of accuracy and precision are obtained by SVM, while the NB has not succeeded to outperform the other classifiers, despite its effectiveness in other classification applications, for any of the topics. This is probably because the NB requires the use of more detailed features.

Several recent research papers have been developed in online recruitment fraud detection. Ref. [11], the authors develop an intelligent online recruitment fraud detection model. The proposed model is based on machine learning and classification techniques. The model was applied to the EMSCAD dataset. The authors applied several algorithms for classification, such as Naive Bayes, zeroR, Logistic regression, decision tree, oneR, and random forest. They used some characteristics of the dataset, such as company profile, industry, and company logo, for classification. The proposed model achieved 97.41% accuracy when using random forest classification. At the same time, naïve Bayes and decision tree gained 38% and 84.77%, respectively. Finally, logistic regression, OneR, and ZeroR achieved 77.22%, 77.33%, and 50%, respectively. This paper did not work on the company profile, requirements, or features to reach higher results.

In Ref. [12], the authors developed an automatic model to detect online recruitment frauds using machine learning techniques. The proposed model is applied to the EMSCAD dataset. The author chose a random group corresponding to the number of real jobs and the number of fraudulent jobs, which are 450 ads. The authors also applied several classification algorithms, such as Random Forest, Naïve Bayes, decision tree, logistic regression, oneR[13], and zeroR[14] classification. The author used some advanced characteristics of the dataset such as job description and benefits in addition to the company profile. The proposed model achieved (90.56 %) for random forest technique, 88.11% for Naïve Bayes, 90.65% for the decision tree, 90% for Logistic regression, 84.33% for OneR, and 50% for ZeroR.

This paper relied on studying the main characteristics of the job, such as job title, location telecommuting, and company logo. While the company's data was neglected, it hopes in the future that there is a relationship between the nature of the job and the company's specifications and its website, to develop a tool capable of detecting job fraud on the Internet.

In Ref. [15], the authors used 21 characteristics of the EMSCAD dataset, which are divided into three descriptive, contextual, and linguistic data. The contextual data was related to the job, while the linguistic data was related to the employer. In addition, the metadata was related to the workplace and the company logo.

The authors developed a learning model and applied three classification techniques which are logistic regression, decision tree, and random forest. It is based on extracting unwanted words in the company profile, benefits, and so on. Then, the feature vector is given the ORFDetector, which is Ensemble Learning Based Online Recruitment Fraud Detection to predict the reality of the job. The proposed model achieved similar results among the three classifications. Where logistic regression, decision tree, and random forest achieved 95.30%, 95.50%, and 95.50% respectively

In Ref. [16], the authors suggested a model for detecting fraudulent jobs using machine learning algorithms, the model was applied to EMSCAD dataset, in addition to new features were extracted from the dataset such as the "Has_Multiple_Jobs" feature. Sometimes an organization posts multiple jobs on a similar field and the "Title_has_extra_info" feature. The purpose of this feature is to examine whether the job title has additional information, such as years of experience, and industry, and ensure the feature selection process, for any classification problem that will help improve classification accuracy using the appropriate machine learning model. The author applied seven algorithms for classification, such as Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), Multilayer Perceptron neural network (MLP), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). The proposed model achieved (97.94%) as an accuracy when using XGBoost (XGB) classification, Naïve Bayes (81.25%), SVM (83.42%), MLP (87.28%), DT (89.42%), KNN (90.02%) and Random Forest (93.83 %).

Finally, we discussed five recent research on job fraud detection. The researchers used many classification algorithms; the highest accuracy in [10] was for the random forest classifier, with an accuracy of 95.35%, while in [11], the highest accuracy was for the random forest classifier, with an accuracy of 97.41%. Vidros and Kolias *et al.* [12] used the same algorithms as in [11], but with a difference in the number of data points used. Where only 450 ads were used, the results of his research showed that the highest accuracy was in the random forest classifier, with an accuracy of 95.5%. Lal and Rishabh *et al.* [15] used 21 features and three classifiers, and the accuracy was very similar, up to 95.50%. Ref. [16], seven classifiers were applied, with the highest XGBoost classifier accuracy at 97.94%.

III. Proposed Model

The key objective of the proposed model is to develop an Arabic fraudulent job detection model that can detect fraud from online recruitment posts. To achieve this objective, the research work focused on achieving the following:

- Translating the unique existing EMSCAD dataset for online recruitment from the English version to the Arabic version.
- Preprocessing the EMSCAD dataset to enhance the performance and accuracy of the proposed model.
- Weighting and determining the most relevant features used to detect fraud in the recruitment process.
- Building the Arabic classification model using the best classifier algorithm.

The proposed model was applied and tested on the Employment Scam Aegean Dataset (EMSCAD) dataset, compiled from real life, and is freely available for researchers. It is a collection of structured and unstructured data. For example, each field may be a string, nominal, and binary data type.

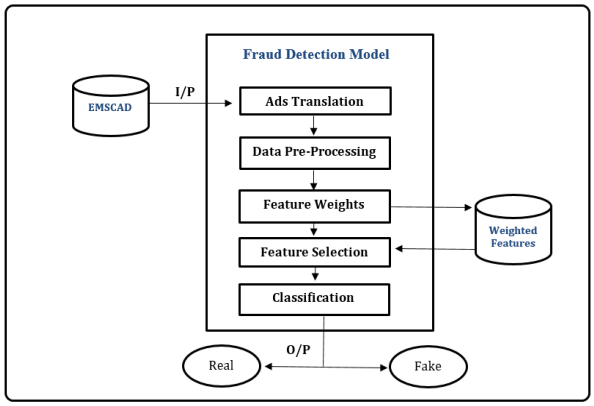


Figure 1 The proposed architecture

It consists of 16 features. These features are divided into three sections. Firstly, features related to the type of employers, such as task description and the nature of the industry. Secondly, features related to the Job description include the number of experience years and level of

education. Finally, features related to the company, such as the company profile, location, and interview questions.

In Fig. 1, the proposed model consists of five phases: data translation, data preprocessing, feature weights, feature selections, and applying classification algorithms.

A. Translation

Our study is focused on the Arabic language, and the available dataset is in the English version. Further, there is a lack of an Arabic dataset. So, there is a need to translate the available English EMSCAD Dataset into Arabic. A googletrans library was used to translate the text from English to Arabic. This step is the basis of the model to identify fraudulent job advertisements in Arabic. Table shows a sample of the translated EMSCAD. There are many translators, Google Translate API and Translator IO. But Google Translator supports the Arabic language strongly, and it is the most used, in addition to being free, unlike translator IO. The accuracy of the Google translation API varies from language to language, as a translation from English to Spanish or French reaches 90%. While it can be less for other languages, such as Arabic, because it depends on the word's literal meaning, Translation saves time and effort. Therefore, it works as an aid but will never replace human translators. The google translator API has probably improved by about 34% using the 51 languages available eight years [17].

B. Preprocessing

It is one of the most critical steps in the proposed model. Due to the nature of the dataset, it contains some data that affect the quality and performance of the model. It includes the following:

- Cleaning data from noise and unwanted characters, such as undefined characters (Â, ©, /) and HTML tags, in addition to the duplicate records.
- Completing incomplete data. Such as some salaries, job benefits, industry, company profile, job description, and requirements.
- Transforming the data into the appropriate form for the classification model. Such as (Education level, location, industry, and company profile) dealing with continuous variables. Some of them include converting.

TABLE I. DATASET TRANSLATION

Feature Name	English Text	Equivalent Arabic Text
Company Profile	Valor Services provides Workforce Solutions that meet the needs companies across the Private Sector, with a special focus on the Oil & Gas Industry. Valor Services will be involved with you throughout every step of the hiring process and remain in contact with you all the w through the final step of signing of the employment	تقدم خدمات فولر حلول القوى العاملة التي تلبي احتياجات الشركات عبر القطاع الخاص ، مع التركيز بشكل خاص على النفط وأمبير. صناعة الغاز. ستشارك خدمات فولر معك طوال كل خطوة من خطوات عملية التوظيف وستبقى على اتصال بك طوال الطريق من خلال الخطوة الأخيرة لتوقيع عقد العمل مع صاحب العمل الجديد.
Description	Our client, located in Houston, is actively seeking an experienc Commissioning Machinery Assistant that possesses strong supervis skills and has an attention to detail. A strong dedication to safety is a m The ideal candidate will execute all activities while complying with qualit requirements and health, environmental, and safety regulations.	يسعى عميلنا ، الموجود في هيوستن ، بنشاط للحصول على مساعد آلات التكليف ذو الخبرة والذي يمتلك مهارات إشرافية قوية ويهتم بالتفاصيل. التفاني القوي في السلامة أمر لا بد منه. المرشح المثالي سينفذ جميع الأنشطة مع الامتثال لمتطلبات الجودة ولوائح الصحة والبيئة والسلامة.
Benefits	In return we'll pay you well, give you some ownership in the compan (stock options) and importantly provide you with excellent opportunities f D G Y D Q F H P H Q W D Q G S U R I H V V L R Q D O G Y H O R S P H Q W R S S R U W X Q L W L H V new pair of Adidas trainers when you join.	في المقابل ، سوف ندفع لك جيدًا ، ونمنحك بعض الملكية في الشركة (خيارات الأسهم) ونوفر لك بشكل مهم فرصًا ممتازة للتقدم والتطوير المهني. أوه ، وسنمنحك زوجًا جديدًا من أحذية Adidas الرياضية عند الانضمام.
Requirements	Implement pre-commissioning and commissioning procedures for rotat equipment. Execute all activities with subcontractor's assigned crew that pertains to the discipline. Ensure effective utilization of commissioni manpower and consumables. Ensure the execution of vendor specialist's field activities with the assigned resources from the sub-contractor per vendor's representative plans. Carry out equipment inspections w client representatives and ensure proper certification is produced. activities. Prepare forms for all pending tests and submit signed certifica for final hand over to the certification engineer for QA and QC	تنفيذ إجراءات التكليف المسبق والتكليف بالمعدات الدوارة. تنفيذ جميع الأنشطة مع الطاقم المعين من قبل المقاول من الباطن والذي يتعلق بالانضباط. ضمان الاستخدام الفعال لتكليف القوى العاملة والمواد الاستهلاكية. المقاول من الباطن لكل خطط تمثيلية للبائع. قم بإجراء عمليات تفتيش على المعدات مع ممثلي العملاء وتأكد من إنتاج الشهادة المناسبة. أعد النماذج لجميع الاختبارات المعلقة وقدم الشهادات الموقعة للتسليم النهائي إلى مهندس الشهادات لضمان الجودة ومراقبة الجودة.

TABLE II. DATA CLEANING SAMPLES

Text before Cleaning	Text after Cleaning
الفوائد الخاصة بك أن تكون جزءًا من شركة سريعة النمو في صناعةمزدهرة اتخاذ قرارات سريعة بفضل التسلسل الهرميالمسطح والهياكل الواضحة الحرية في الكشف عن أفكارك الخاصة & amp ؛ الإبداع وتحمل المسؤولية منذ البداية استمرار النمو في فريق دولي ناجح يزدهر في جو عمل مألوف ولكن احترافي ومشروبات مجانية وتنسالطاولة وتناول طعام الغداء وإدارة تشعر بالسعادة وفعاليات الفريق Your Benefits Being part of a fast growing company in a thriving industry Quick decision making thanks to a flat hierarchy and c structures Freedom to reveal your own ideas & Creativity accountability from the start Continuing to grow into a succes international team that thrives in a familiar yet professional wor atmosphere, free drinks, table tennis, lunch, happy managemer team events	الفوائد الخاصة بك أن تكون جزءًا من شركة سريعة النمو في صناعة مزدهرة اتخاذ قرارات سريعة بفضل التسلسل الهرمي المسطح والهياكل الواضحة الحرية في الكشف عن أفكارك الخاصة الإبداع وتحمل المسؤولية منذ البداية استمرار النمو في فريق دولي ناجح يزدهر في جو عمل مألوف ولكن احترافي ومشروبات مجانية وتنس الطاولة وتناول طعام الغداء وإدارة تشعر بالسعادة وفعاليات الفريق Your benefits Being part of a fast-growing company in a thriving industr Quick decisions thanks to a flat hierarchy and clear structures Freed reveal your own ideas Creativity and accountability from the s Continued growth into a successful international team thriving in a familia yet professional work atmosphere Complimentary drinks, table tennis dining Lunch and management feel happy and team events

1) Data cleaning

In this step, we understand the data well to study its nature. Then, the noise letters and words from the text are removed. In addition, the data may contain Latin letters, HTML tags, and misspellings during the entry process. This affects the quality of the data. Therefore, we need to identify and remove unwanted characters. Furthermore, the repeated words from the text are removed. Table I shows a sample of the data cleaning.

2) Completing missing data

EMSCAD dataset contains missing values for several reasons, such as notes that were not recorded, data corruption, input error, and many other reasons. Processing missing data is essential because many machine learning algorithms do not support data with missing or incorrect values that can affect the quality of the results. The database contains different data types: binary, String, and Nominal. Specifically, there are missing notes for some columns marked as zero. We can confirm this by defining those columns and knowing the field that the zero value is not valid for those metrics such as salary, company file and benefits, and personal interviews. As a result, correct knowledge patterns are discovered, and precise decision-making based on the merger of diverse algorithms is obtained. It assures that accurate data cleansing methods are employed, and it cleans Arabic datasets using a multi-level cleaning process based on Arabic Misspelling Detection and Correction Model (AMDCM), and Decision Tree Induction (DTI) [18]. Fig. 2 shows the Arabic misspelling and correction model such as والفاعليين´3 will after correcting والفاعلين´3 shown in Table III .

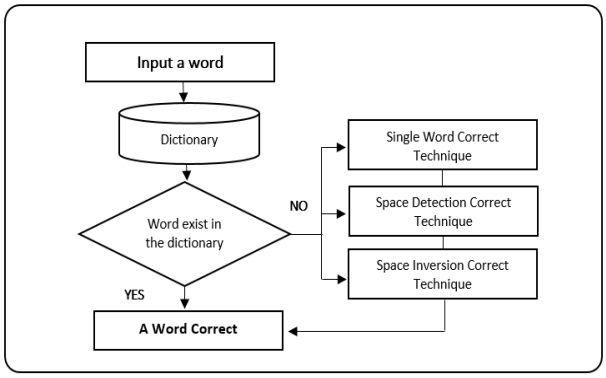


Figure2. Arabic misspelling detection and correction model

TABLE III. ARABIC WORD MISSPELLING

Misspelling Word	Word Correction
والفاعليين (actors)	والفاعلين(actors)
والةالقهوة (coffee machine)	والة القهوة (coffee machine)
سباقا (racing)	سباقات (racing)
المسطحالهرمي (flat pyramid)	المسطح الهرمي (flat pyramid)

3) Data transformation to algorithm formats

Data may have one of a few types, such as numeric or categorical, with subtypes for each, such as integer and real values for numeric, and nominal, ordinal, and Boolean for categorical. The process of converting text into a suitable method for classification and data cleaning by studying the nature of each Feature.

a) Numerical data

Salary is Quantitative data, the natural way to represent these salary ranges is numerical. Therefore, we need to rescale the values of a numerical feature to be between two values. Typically 0 to 1.

In the EMSCAD dataset the salary feature as range values for example 20000-28000. To normalize salary column we need to calculate average for each salary range ANG = 20000 + 28000 /2 = 24000.

Then apply scaling formula [19, 20]

$$X\ normalize = \frac{X - Xmin}{Xmax - Xmin}$$

Max salary = 762000 & min salary = 0, Assume zero to advertisement with undefined salary

$$X\ normalize = \frac{24000 - 0}{762000 - 0} = 0.0314$$

C. Categrical Data

EMSCAD contains categorical data such as Require Education, Employment Type, and Industry. Therefore, we use dummy variables to represent it. Dummy variables are limited to two specific values, 1 or 0. Typically, 1 represents the presence of a qualitative attribute, and 0 illustrates the absence. The number of dummy variables required to describe a particular categorical variable depends on the number of values the categorical variable can assume.

For example, the Work Experience feature consists of 6 values. We could represent each value with two dummy variables as shown in Table IV .

TABLE IV. CATEGORICAL DATA TRANSFORMATION SAMPLES

Work Experience	Sample 1 (Internship)	Sample 2 (Associate)	Sample 2 (Entry level)
Not Applicable	0	0	0
Internship	1	0	0
Entry level	0	0	1
Mid-Senior level	0	0	0
Associate	0	1	0
Executive	0	0	0

IV. FEATURE WEIGHT

Feature weight is used to approximate the effect score for features. The main objective of using feature weight is to identify the effect of each feature on the accuracy of the model results and reduce the multiplicity and variety of the data and have a more significant impact on the output Where a high value for the effective feature and a value approaching zero for the weak feature as shown in Fig 3.

For feature weight, we used the random forest technique to see how each feature affects the impurity of the split (the feature with the highest decrease is selected for the internal node). We may calculate how much each feature reduces impurity on average; the average number of trees in the forest is used to determine the feature's relevance.

We use Eq. (1) to applying features weights, aggregate the votes (predicted classifications) across trees by applying the weights W_j to the votes. Let $\upsilon_{test,i,j}$ be the vote for tree j for subject i in the independent test data, where i=1, $M2$=N/4 [21].

$$WP_i = \sum_{j=i}^{ntree} \boldsymbol{W_j}.\, \upsilon_{\boldsymbol{test,i,j}}$$

Using Eq. (1), we can compute performance measures of the weighting procedure in the independent test set, such as the Prediction Error (PE_{WRF}) and AUC (AUC_{WRF}) of the Weighted Random Forest. If y_i is the true class of subject i, then the prediction error PE_{WRF} can be computed based on the weighted classification WC_i:

$$WC_i = I(WP_i \geq 0.5)$$

$$PE_{WRF} = \frac{1}{M_2} \sum_{i=1}^{M_2} |WC_i - y_i|$$

After feature weighting, feature selection is applied to determine which features impact recruitment fraud detection most, which plays a key role in the accuracy of the results. According to weights that are generated using the random forest, the top five important weighted features that influence are chosen as shown in Table V.

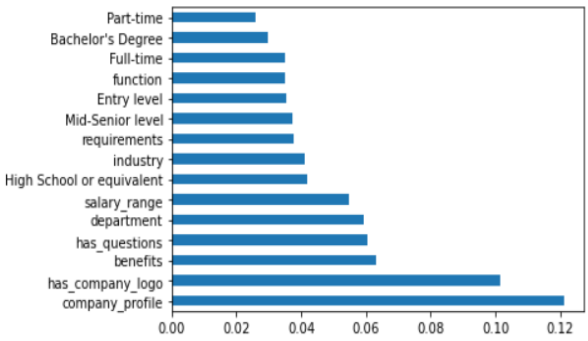


Figure 3 Features weights based on random forest technique

TABLE V. TOP FEATURES WEIGHT

Feature	Weight
Company Profile	12.108663
Has Company Logo	10.137965
Benefits	6.292445
Has Questions	6.046285
Department	5.913834

V. CLASSIFICATION

Use classification techniques to check for fraudulent jobs and track and compete results to determine the best fraud detection model on the internet. Using the necessary parameters, the classifiers are trained. Default parameters may not be sufficient to maximize the performance of these models. The model dependability is improved by adjusting these parameters.

To apply the ensemble classification, the training set and testing set can be defined by specifying the ratio of data, which is 80% training, and the rest is the testing set (20%). The training data set contained 14304 records, while the test dataset contained 3576 records. The cross-validation value specifies the statistical performance of learning and the accuracy of predicting. The proposed model is developed using five machine learning classifiers, which are Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), k-nearest Neighbors (KNN), and Support Vector Machines (SVM).

RF as an ensemble classifier is built based on 100 estimators with 42 random_state, n_jobs with three processors, true value for bootstrap, true value for oob_score, and validating the criterion parameter with entropy. The DT classifier utilized the default parameters with Gini criterion. NB (GaussianNB) classifier utilized the default parameters for var_smoothing of 1e-09. The K-NN classifier gives a promising result for the value k=5 considering all the evaluating metrics. The SVM classifier trained the data with regularization of value of 2 and linear kernel.

A. Performance Evaluation Metrics

When analyzing a model's performance skill, some measures must be used to justify the evaluation. The following metrics are considered to choose the best suitable problem-solving technique. Accuracy in Ref. [22] is a metric that measures the proportion of accurate predictions to the total number of examples taken into account. However, because it does not account for incorrectly anticipated cases, accuracy may not be a sufficient criterion for evaluating a model's performance. A serious problem arises when a bogus job is treated as legitimate. As a result, false positive and false negative scenarios that compensate for misleading must be considered. Precision and recall must be taken into account while measuring this compensation.

The ratio of true positive results to the number of positive results predicted by the classifier is known as precision [22]. The number of correct positive outcomes divided by the total number of relevant samples is called recall [22]. The F1-Score or F-measure [22], which is determined as the harmonic mean of precision and recall, is a parameter that is concerned with both recall and precision.

In addition to fold cross-validation is a standard method for estimating the performance of a machine learning algorithm on a dataset. The estimate of model performance depends on k-fold. The value of k is essential to compare this to an ideal test condition. This can help to choose an appropriate value for k. Once a value is selected, it can be used to evaluate a suite of different algorithms on the dataset. The distribution of results can be compared to an evaluation of the same algorithms using an ideal test condition to see if they are highly correlated or not. If correlated, it confirms the chosen configuration is a robust approximation for the perfect test condition [23].

VI. EXPERIMENTAL RESULTS

Three experiments have been conducted using the translated EMSCAD dataset. Classifiers have been trained and tested to detect job fraud based on the dataset that contains fake and real ads. The data were divided into two

groups. First, the training group was 80% of the data, representing (14304 job ads). At the same time, the second was the test group which represented 20% of the data (3576 job ads).

Dataset We conduct all our experiments on publicly available datasets on the internet (EMSCAD) [9]. There are 17,880 annotated jobs in this dataset, with 866 (4.85%) being fraudulent and 16994 (95.15%) being legal. The ratio of fraudulent versus legitimate jobs in this dataset is substantially skewed. It consists of 16 features, as shown in Table VI. Also, Table VII shows statistics about EMSCAD Dataset. It contains 17,880 job postings, 866 are fraud, and 17,014 are real ads.

These features are divided into three sections. Firstly, features related to the type of employer, such as task description, and the nature of the industry. Secondly, features related to the Job description include the number of experience years and level of education. Finally, features related to the company, such as the company profile, location, and interview questions.

TABLE VI. A DETAILED FEATURES OF EMSCAD DATASET

Feature Name	Data Type	Description	Relevant To
Company Profile	String	A brief description of the company	Company
Location	String	The location of the Company	
Benefits	String	Benefits list offered by Company	
Telecommunication	Binary	True If work from home & False If work from Company	
Company Logo	Binary	True if company logo exists	
Has Questions	Binary	True if an interview question exists	
Title	String	The job advertisement Title	Job
Salary-Range	Numerical	Suggested Salary Range	
Department	String	Job relevant department	
Require Education	Nominal	Doctorate, Master's Degree, Bachelor's, etc.	
Requirements	String	Required list for job	
Employment Type	Nominal	Full-type, Part-time, Contract, etc.	
Require Experience	Nominal	Executive, Entry level, Intern, etc.	
Function	Nominal	Consulting, Engineering, Research, Sales etc	
Description	String	Describe task of Job seeker	Job Seeker
Industry	Nominal	Automotive, IT, Health care, Real estate, etc.	

TABLE VII. STATISTICAL DATA ABOUT EMSCAD

Total job ads	17860
Fraud jobs	866
Real jobs	16994
Period interval	2011-2014

A. *Full Featured Experiment*

In this experiment, we used all Arabic EMSCAD features such as Location, Department, Salary-Range, Company Profile, Description, Requirements, Benefits, Telecommunication, Company Logo, Has Questions, Employment Type, Require Education, Require Experience, Industry, and Function. The dataset is recorded into two groups. First, the training group was 80% of the data, representing (14304 job ads). At the same time, the second was the test group, and represented 20% of the data (3576 job ads).

TABLE VIII. PERFORMANCE MATRIX COMPARISON TABLE FOR CLASSIFIER-BASED PREDICTION AFTER APPLYING ALL FEATURES IN EMSCAD DATASET

Performance Measure Metric	Accuracy	Precision	Recall	F1-Score
RF	0.96	0.83	0.67	0.72
NB	0.15	0.52	0.54	0.14
SVM	0.95	0.87	0.51	0.51
DT	0.95	0.69	0.81	0.73
KNN	0.95	0.80	0.67	0.71

Five classification algorithms were applied: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF). Fig. 4 shows the comparative study of the classifiers concerning evaluation metrics. The results were satisfactory as the Random Forest classifier. It is the highest accuracy among all the classifications, with 96%. While the decision tree achieved 95%, and the support vector machine gained 95%. Finally, the K-Nearest Neighbor reached 95%, and Naive Bayes had a lower accuracy, with 15%. Table VIII shows the Performance Measure Metric by Macro averaging for the proposed classifiers model.

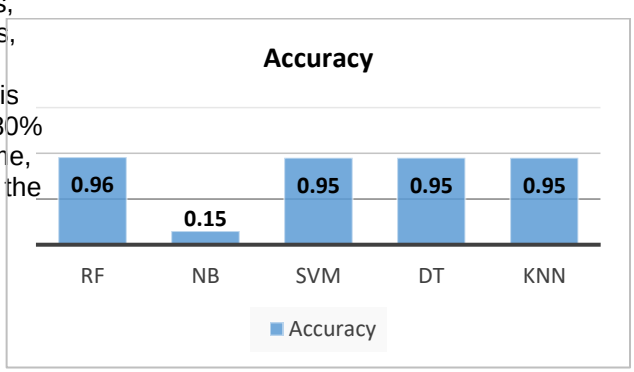


Figure 4. Performance comparison chart for classifier based prediction after applying all features in the EMSCAD dataset

B. *Weighted Featured Experiment*

The feature weight algorithm for all dataset features is applied. It is important to study what features significantly impacts the results. Fig. 3 shows the order of the features according to the degree of their effect on the nature of the job, fake or trust are: The highest priority was the company profile feature, has company logo feature, has questions

feature, benefits feature, department feature, salary range feature, high school feature, and industry. Then, the feature selection algorithm is applied to choose the essential feature. The top five features, with more than 50%, are preferred. According to Fig 3, the selected feature is company profile, has company logo, has questions, benefits, and department feature. Fig 5 shows a significant improvement using the naïve Bayes classifier with 88% accuracy compared with the previous Full Featured Experiment. While the SVM, DT, and RF achieved the same result, 95%, 95%, and 96%, respectively but the accuracy of the KNN classifier decreased to 94%. Table IX shows the Performance Measure Metric by Macro averaging for the proposed classifiers model after applying features weight and features selection.

TABLE IX. AFTER APPLYING FEATURES WEIGHT AND FEATURES SELECTION, THE PERFORMANCE MATRIX COMPARISON TABLE FOR ALL SPECIFIED SUPERVISED MACHINE LEARNING MODELS

Performance Measure Metric	Accuracy	Precision	Recall	F1-Score
RF	0.96	0.48	0.50	0.49
NB	0.88	0.60	0.75	0.63
SVM	0.95	0.47	0.50	0.49
DT	0.95	0.50	0.47	0.49
KNN	0.94	0.64	0.55	0.57

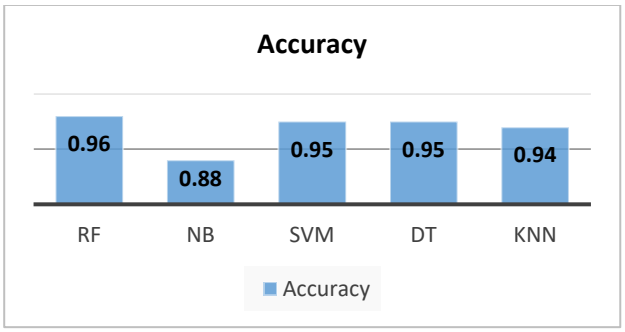


Figure 5 Performance comparison chart for classifier based prediction after applying features weight and features selection

The results of this experiment indicate an increase in the accuracy of NB to 88% and KNN decrease to 94% because we are working on the highest weighted features in the dataset, as shown in Fig 3. Thus, the ratio of correlations between features decreased, so the probability ratio decreased, and the model's accuracy increased. While the accuracy of the other models (SVM, DT, and RF) has not been changed compared to the first experiment due to its dependence on the decision tree.

C. *Benefit-Processing Feature Experiment*

This Experiment is the third one. It depends on processing the text of the benefit feature. Depending on the features resulting from the features weighting processing phase, the features that have the highest weight score: Company Profile, Has Company Logo, Has Questions, Benefits, and Department.

Using Collection-Independent Automatic Keyword Extractor (YAKE) [24]. YAKE is a novel feature-based system for multilingual keyword extraction from single documents. It supports texts of different sizes, domains, or languages. YAKE consists of six main concepts.

- Text preprocessing
- Feature extraction
- Termsscore
- Creating a keyword list of potential candidates
- Data Deduplication
- Ranking

YAKE is used to extract the positive keywords from the benefits feature text of the job. A sample of the extracted keywords from the job benefits feature is shown in Table X.

TABLE X. POSITIVE KEYWORDS EXTRACTED FROM BENEFITS FEATURE

Arabic extracted keywords	Equivalent English extracted keywords
الراتب التنافسي ، الخبرة ، حزمة المزايا الشاملة ، التدريب ، الإجازات ، الوقت المدفوع ، التعلم ، النمو السريع ، التأمين الصحي ، إجازة مدفوعة ، المكافآت ، عمولة ، الأجر ، مكان عمل خالٍ ، مؤسسة أوروبية كبرى ، من التبغ ، حزمة تعويضات تنافسية ، مقهى ستاربكس ، أجر مضاعف ، فرصة سفر دولية	Experience , Competitive Salary , Training , Learning , Paid Time , Vacations , Rapid Growth , Health Insurance , Paid Leave , Commission , Bonuses , Major European Institution , Tobacco Free Workplace , Competitive Compensation Package , Starbucks , Double Pay , Travel Opportunity international

TABLE XI. AFTER EXTRACTING BENEFITS FEATURES KEYWORDS, THE PERFORMANCE MATRIX COMPARISON TABLE FOR ALL SPECIFIED SUPERVISED MACHINE LEARNING MODELS

Performance Measure Metric	Accuracy	Precision	Recall	F1-Score
RF	0.97	0.85	0.75	0.79
NB	0.20	0.53	0.56	0.21
SVM	0.96	0.87	0.51	0.51
DT	0.96	0.75	0.81	0.78
KNN	0.96	0.81	0.71	0.75

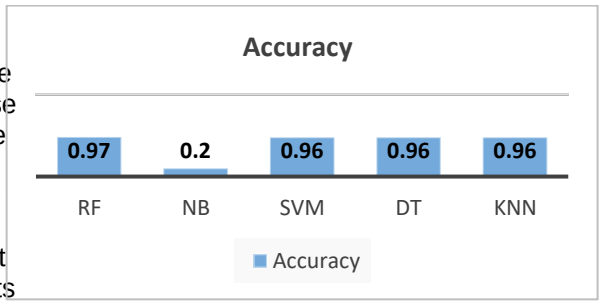


Figure 6 Performance comparison chart for classification after extracting benefits features keywords

After that, the model is provided with data extracted from the benefits feature. The results that the algorithms have achieved are improved, where the random forest classification was performed with the highest accuracy with 97%, decision tree classifier with 96%, Support vector machine classifier with 96%, K-Nearest Neighbor classifier with 96%, and Naive Bayes classifier with 20% as shown in Fig 6. Further, Table XI shows the Performance Measure Metric by Macro averaging for all

supervised machine learning models after extracting benefits feature keywords.

VII. Discussion

The results showed a clear improvement in the dimension of the models used. In the first experiment that depended on all Arabic features according to the dataset, the RF classifier, KNN classifier, DT Classifier, and SVM Classifier achieved a high accuracy with 96%, 95%, 95%, and 95%, respectively. In comparison, the naive Bayes Classifier(NB) achieved 15%, which is very low accuracy. This is due to the nature of the NB classifier, and the features multiplicity used in this model reaches 16 features where it works on probabilities. The correlations between the features are substantial, so the accuracy of this model decreases. Further, the achieved accuracy of the random forest classifier is high because it depends heavily on the decision tree.

In the second experiment that depended on top features weight, the NB classifier achieved 88%, which has improved compared with the accuracy result in the first experiment. This is because the number of used features has been reduced to only 5 features. They are the top 5 weighted features. This reflects the validity of our experiment. If the accuracy of the correlations between features increases, the accuracy of the model increases.

While the KNN algorithm uses K as the nearest neighbors of a given query point, the distance between the (fraudulent) query point and other data points must be calculated. These distance measures help form decision boundaries, but we only used the top five features, so the distance between the point and its nearest neighbor increased. Thus Classification accuracy decreased by 1%. Finally, in the third experiment that depended on the text processing text of benefits feature. The accuracy of the model improved significantly. The RF classifier achieved 97%, the NB classifier achieved 20%, the SVM classifier achieved 96%, the DT classifier achieved 96%, and the KNN classifier achieved 96%. The accuracy results improved by 1 percent over the second experiment within RF, SVM, and DT classifiers while improving by 2 percent in the KNN classifier. Fig 7 shows the accurate comparison between the three experiments results.

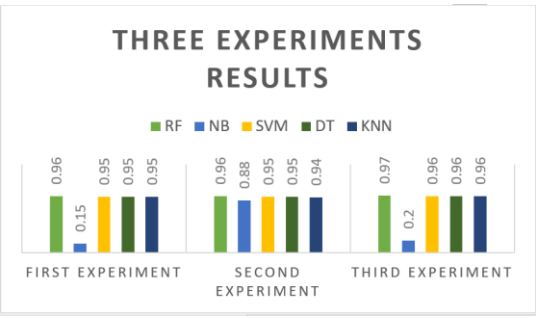


Figure 7. Comparison between three experiments results

VIII. Conclusion and Future Work

Online recruitment systems are a new, easy, and effective way to select the best and most promising method for many companies. Moreover, it expresses a broad vision of the world in 2030, and the percentage of use of these platforms will gradually increase, despite their misuse by criminals and fraudsters and the violation of the privacy of job seekers.

Research papers have studied the problem of reducing job fraud via the Internet. It proposed a model to identify job advertisements as real or fraudulent, which was applied to a data set in the English language.

This research paper proposed a model for detecting job fraud by applying it to the Arabic language. It is considered the first that used the Arabic language. We applied five algorithms, Support Vector Machine(SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and k Nearest Neighbors algorithm (KNN), as ensemble classifiers.

The model achieved 96% accuracy, which is the best result that has been obtained until now, applied to the Arabic language dataset, while the model achieved 97% accuracy applied to the features that have higher weights. As shown in Table V, the results confirmed that the job depends on several features to reveal its identity, such as "company logo", "company profile", "interview questions", "benefits", and "department".

This research used an (EMSCAD) dataset. It is a free dataset available on the Internet. Then it was translated into Arabic.

In future work, we can take advantage of the available data set to examine and analyze new features such as the company profile, know the expected salary, and apply different algorithms. We will try to collect an Arabic dataset. Then, applying the proposed model to the collected dataset. Finally, comparing the results with the results of the translated EMSCAD dataset.

We will work on solving the problem of an unbalanced dataset by using the oversampling method. This method is applied by adding records to the minority class or deleting ones from the majority class. To bring the number of fraudulent ads near the number of actual ads, giving more learning for the model and obtaining higher scores in different evaluation metric [25]. Our goal is to propose an applicable Arabic employment fraud detection tool for commercial.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Author Contributions

Mohamed A. Sofy conducted the research by defining research frameworks, designing research methodology, analyzing, modeling, and writing the paper. Rasha M. Badry conducted the research by reformulating and revising the research paper to become publishable. Mohamed H. Khafagy supported and provided necessary feedback and ideas for the entire research paper. All authors had approved the final version.

REFERENCES

[1] M. Syed, P. Eric, and Kayes, "Online recruitment fraud detection: A study on contextual features in Australian job industries," *IEEE Access*, vol. 10, no. 1, 2022.

[2] S. Morgan, "Cybersecurity ventures," Cybercrime Report, 2017.

[3] A. Timothy, "Random forests, decision trees, and categorical predictors: The absent levels problem," *J. Mach. Learn. Res.*, vol. 19, 2018.

[4] N. Mohapatra, K. Shreya, and C. Ayes, "Optimization of the random forest algorithm," in *Advances in Data Science and Management: Proceedings of ICDSM 2019*, Springer, 2020, p. 1.

[5] W. Aji, K. Ahmad, M. Della, *et al.*, "Naïve bayes classifier for journal quartile classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, pp. 21–26, 2019.

[6] L. Wang, "Research and implementation of machine learning classifier based on KNN," *IOP Conference Series: Materials Science and Engineering*, pp. 10–12, 2019.

[7] F. R. Lumbanraja, E. Fitri, Ardiansyah, *et al.*, "Abstract classification using support vector machine algorithm," *Journal of Physics: Conference Series*, 2021.

[8] G. Babacar, D. Z. Zhang, and W. Aziguli, "Improvement of support vector machine algorithm in big data background," *Mathematical Problems in Engineering*, pp. 1–9, 2021.

[9] Employment Scam Aegean Dataset (EMSCAD) [Online]. Available: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction?resource=download

[10] M. Alkhair, K. Meftouh, N. Othman, *et al.*, "An Arabic corpus of fake news: Collection, analysis and classification," *Communications in Computer and Information Science*, vol. 1108, 2019.

[11] B. Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection," *Journal of Information Security*, vol. 10, July 2019.

[12] S. Vidros, C. Kolias, G. Kambourakis, *et al.*, "Automatic detection of online recruitment frauds: characteristics, methods, and a public dataset," *Future Internet*, p. 3, 2017.

[13] F. Alam and S. Pachauri, "Comparative study of J 48, Naive Bayes and OneR classification technique for credit card fraud detection using WEKA," *Computational Sciences and Technology*, vol. 10, June 2017.

[14] C. L. Devasena, "Effectiveness analysis of ZeroR, RIDOR and PART classifiers for credit risk appraisal," *International Journal of Advances in Computer Science and Technology*, pp. 6–11, 2014.

[15] S. Lal, J. Rishabh, S. Neetu, *et al.*, "ORFDetector: Ensemble learning based online recruitment fraud detection," in *Proc. Twelfth International Conference on Contemporary Computing (IC3)*, 2019.

[16] A. Mehboob and M. S. I. Malik, "Smart fraud detection framework for job recruitments," *Arabian Journal for Science and Engineering*, 2021.

[17] Aiken and Milam, "An updated evaluation of google translate accuracy," *Studies in Linguistics and Literature*, 2019.

[18] [S. Khaled, R. Aref, and A. Fahmy, "An approach for analyzing and correcting spelling errors for non-native Arabic learners," in *Proc. the 7th International Conference on Informatics and Systems*, 2010.

[19] M. Civera and C. Surace, "Imbalanced multi-class classification of structural damage in a wind turbine foundation," *Lecture Notes in Civil Engineering*, vol. 2706, 2022.

[20] C. X. Hang, S. Ivan, and O. Zoran, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinformatics*, p. 9, 2016.

[21] K. Gajowniczek, I. Grzegorczyk, T. = E N R, *Zev*, N, L, "Weighted random forests to improve arrhythmia classification," *Computational Intelligence for Physiological Sensors and Body Sensor Networks*, pp. 1–8, 2020.

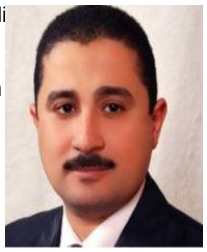
[22] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, March 2015.

[23] D Berrar, *Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, January 2018.

[24] R. Campos, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, January 2020.

[25] M. Roweida, R. Jumanah, and A. Malak, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th International Conference on Information and Communication Systems (ICICS)*, 2020.

Mohamed A. Sofy is currently a teaching assistant at the Faculty of Computers and Information, Fayoum University. He received a B.S. degree from the Faculty of Computers and Information Fayoum University, Egypt, in 2019.
He is currently pursuing a master's degree from the Faculty of Computers and Information, Fayoum University, Egypt. His research interests include machine learning, data mining, nlp, and database.



Mohamed H. Khafagy is the head of Big Data Research Group at Fayoum University. Mohamed received his Ph.D. in computer science in 2009. He also works at Oracle Egypt as a consultant. Mohamed is the manager of the National Electronic Exam center in the Supreme Council of Universities. Mohamed worked as a postdoc in the DIMA group at Technique University Berlin in 2012. Mohamed established the first big data research group in Fayoum University in 2013. He has many publications in the area of big data, cloud computing, and database.

Rasha M. Badry is a lecturer at Information Systems Department, Faculty of Computers and Information, Fayoum University, Egypt. She is director of Crisis Management, Faculty of Computers and Information, Fayoum University. She also director of National Bank for Scientific Laboratories and Equipment, Supreme Council of Universities, Egypt. She got her Ph.D., M.Sc., and B.Sc. in 2015, 2007, and 2003 from Faculty of Computers and Information, Helwan University, Egypt. Her research interests are NLP, machine learning, data science.