Performance Evaluation of Different Data Mining Techniques for Social Media News Credibility Assessment

Sahar F. Sabbeh[1, 2]

College of computer science and engineering, University of Jeddah, KSA[1]
Faculty of computing and information sciences, Banha University, Egypt[2]

***Abstract*—Social media has recently become a basic source for news consumption and sharing among millions of users. Social media platforms enable users to publish and share their own generated content with little or no restrictions. However, this gives an opportunity for the spread of inaccurate or misleading content, which can badly affect users' beliefs and decisions. This is why credibility assessment of social media content has recently received tremendous attention. The majority of the studies in the literature focused on identifying features that provide a high predictive power when fed to data mining models and select the model with the highest predictive performance given those features. Results of these studies are conflicting regarding the best model. Additionally, they disregarded the fact that real-time credibility assessment is needed and thus time and resources consumption is crucial for model selection. This study tries to fill this gap by investigating the performance of different data mining techniques for credibility assessments in terms of both functional and operational characteristics for a balanced evaluation that considers both model performance and interoperability.**

***Keywords*—Data mining; performance evaluation; news credibility; Twitter; social media**

I. Introduction

Social media platforms suffer from the lack of supervision over content which can result in the spread of inaccurate (fake) information either unintentionally or intentionally for deceptive purposes. That is why using data mining models for content credibility assessment has become an important practice in the context of social media. To date, the bulk of the work in the literature focused on identifying the most informative features for higher precision credibility. Features are extracted at different levels (user, topic and propagation) level [1]-[9]. User-related features such as account/profile, demographics, age, account age, followers, photo, behavior (tweeting, retweeting) can be extracted and used to evaluate source credibility as inaccurate news can probably be created and spread by automated software agents or fake accounts created only for this sake. Topic related features can content related or contextual related. Content - related features include visual and textual features which can be collected and analyzed using standard NLP and text analysis techniques (i.e. images included, Hash-tags, URLs, sentiments/subjective content, etc.). Contextual information includes topic headlines, users' comments, rates, likes, emotional reactions, number of shares, etc. For example, topic/post headline may be misleading (known as "clickbait") which implies non-credible content or at least irrelevant content.

The extracted features are fed into different classification models which are then evaluated to identify the best performance given those set of features. The used techniques include: Logistic Regression (LR) [9]-[12], Decision Trees [1], [3], [6], [13]-[17], [19], Artificial Neural Networks [20],[21],[22], Support Vector Machines(SVM) [6], [13], [14], [15], [17]-[21], [23] Random Forest (RF) [13], [15], [18], [24], Naïve Bayesian (NB) [6], [16]-[19], [21] and K-nearest Neighbor (KNN) [17], [20], [21]. SVM and Decision Trees are the most known and widely used models. Very few works tried to use other models such as Linear Discriminant analysis (LDA) [21] and Adaptive Boosting (Adaboost) [23]. The performance of data mining techniques for credibility analysis included only the most well – known techniques disregarding more advanced techniques that may better utilize the extracted features such as bagged and boosted ensemble models. Moreover, the results of the performances are difficult to compare as each study recommends a different model and therefore, no general agreement can be reached.

Additionally, those studies focused only on the functional capabilities of the models by evaluating their predictive power, which, despite being important, is not enough. Operational characteristics are as important as functional capabilities. These include evaluating time and memory usage during both training and runtime. That is, measuring the amount of time and memory during model training and the amount needed to classify new data. As long classification time or excessive memory usage may mean that the model is unsuitable for real-time environments. Thus, a benchmark or empirical analysis that focuses only on the predictive performance will be insufficient to evaluate models operability.

This study tries to fill the gap in the research by focusing on news credibility assessment on Twitter as a case study. The used dataset for this study is publically available at GitHub1, it contains a set of 9252 Twitter news related to US election 2016 represented by 23 mixed features (numerical and binary). A set of 13 chosen models that represent different learning models were used in this study to provide an empirical analysis of different models and to identify the extent to which they are applicable for credibility assessment. LDA was selected as a

[1] https://github.com/marianlonga/FakeNews

linear learning model, mixture discriminant analysis (MDA), SVM, KNN, and NB. Both Multi-layer perceptron (MLP) and learning vector quantization (LVQ) were selected as ANNs. CART and C50 represent tree-based models and finally, Bagging CART (BaggedCart), ADAboost, Gradient boosted machine (GBM) and RF represent ensemble learning models. The selected models are evaluated based on accuracy, precision, recall, F-measure and computational time (processing and classification) and memory usage.

This paper is organized as follows: in Section II, a review of the previous empirical analysis of different data mining models in credibility assessment is presented. Section III provides a step-by-step description of the study methodology. Experimental results are discussed in Section IV. And Section V concludes the study and sheds light on study limitations and possibilities for future work.

II. Data Mining for Credibility Assessment in Social Media

Data mining is a process that aims to analyze, identify hidden patterns, and discover knowledge from large volumes of data. Classification techniques are supervised techniques that classify data item into predetermined classes. These techniques construct models using the labeled data to predict the label of unknown data sets.

The data mining process begins by applying data preprocessing (i.e. data transformation, cleaning, feature selection, etc.) is applied to improve the classification efficiency of the algorithm. The data set contains each tuple is labeled to belong to a predefined class. Part of the tuples is used for model construction (training dataset). The models are represented as classification rules or mathematical formulae and are tested using a set of independent data samples/tuples (test dataset) otherwise overfitting may occur. Finally, accuracy rate of the model is calculated as the percentage of test set tuples that are correctly classified by the model. Data mining techniques have been used for assessing the credibility of both information content and source. Credibility is assessed in terms of multiple features that are related to the news source, content and propagation medium. Data mining techniques use the features at one or more levels to label information content and/or source as credible/non-credible or fake/real. The comparisons summarized in Table I were performed among different models have conflicting results regarding their relative performance to one another. In the work [6], [19], DT achieved higher performance than SVM while in [15] SVM achieves better performance than DT. In [17], two different datasets were used and DT achieved the highest performance among other models given the first dataset while KNN was the best given the second dataset. In [20] LR model outperformed more sophisticated non-linear models such as ANN, DT, and SVM. However, ANN proved higher performance in [21]. Ensemble models RF in [13], [18] and Adaboost in [23] proved higher performance over SVM.

In conclusion, there is a need for a unified study that analyzes the performance of different models and evaluates their performance and applicability for credibility assessment.

TABLE. I. Summary of Empirical Studies of Data Mining Models for Credibility Assessment

Study	Models	Best performance
[1]	• (SVM) • Decision trees • extremely randomized trees (ERT) • Naive bayes	**ERT**
[6]	• SVM • Decision trees • Bayes networks	**Decision tree**
[13]	• Decision trees • Random Forest • SVM	**Random Forest**
[14]	• Decision trees • SVM	**Decision tree**
[15]	• Decision tree • SVM • Random Forest	**SVM**
[16]	• Decision tree • Naïve Bayes	**Decision tree**
[17]	• SVM • Naïve Bayes • KNN • decision trees	• **Decision tree for 1st dataset** • **KNN for the 2nd dataset**
[18]	• Naïve Bayesian • SVM • Random forest	**Random Forest**
[19]	• Decision tree • Naïve Bayesian • SVM	**Decision tree**
[20]	• Logistic Regression (LOG) • SVM • KNN • ANN • Decision trees	**Logistic Regression**
[21]	• ANN • KNN • SVM • Naive Bayes • Linear discriminant analysis (LDA)	**ANN**
[23]	• SVM • Adaboost	**Adaboost**

III. Methodology

A. Dataset

The used dataset contains twitter news related to US elections 2016. The dataset contains 9252 Twitter news represented by 22 explanatory variables and one response variable. The predictors are related to both news content and source. The target variable labels each tweet to be fake/non-fake represented by (True/False) variable. The dataset contains 254 instances labeled “unknown” and 2749 with no label. For this study unlabeled observations and noisy/unknown ones were disregarded.[2] The remainder 5598 include approximately 87% labeled false/ to indicate non-fake/real news or other type of news (i.e. comment, etc.), where 13% are labeled True to indicate fake news. Dataset metadata is presented in Table II.

[2] Dealing with missing and noisy labels are out of the scope of this study.

B. Data Preprocessing

1) Data transformation

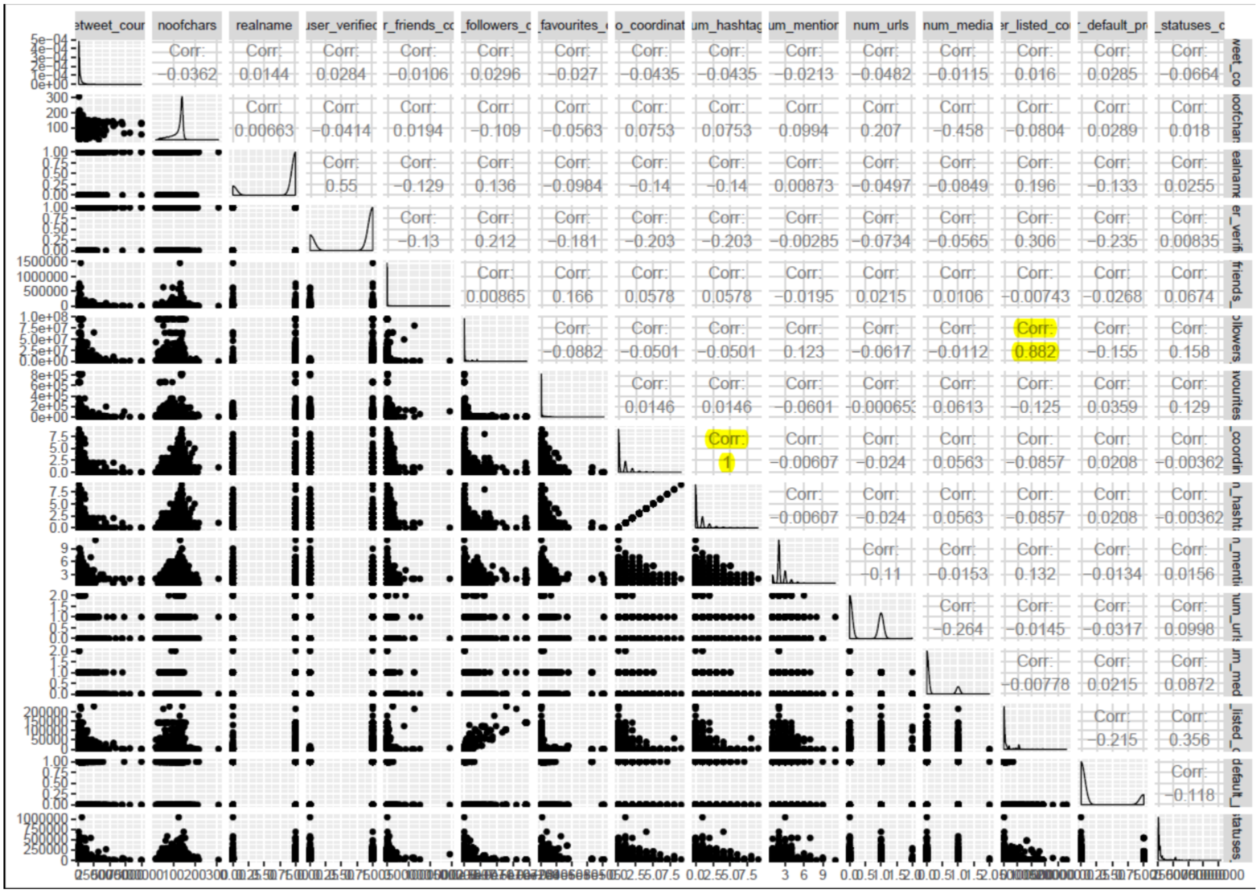
- The variable “Tweet Id” was removed for its irrelevance to the problem.
- The variable “text” was used to derive a new variable “noofchars” to indicate the number of characters in each tweet.
- The “Description” variable was removed and instead, a Boolean variable was added to indicate whether or not user profile has a description.
- “text”, “description”, “Tweet_id”, “created_at” and “tweet_source” were removed from the dataset as they are considered unrelated to the classification problem.
- The variables “user_name” and “user_screen_name” were removed and instead a new derived variable that indicates whether or not the account has a real/nick name is added.
- Binominal variables (user_verified, isfake with (true/false) values was transformed into binary (0/1).

2) Explanatory data analysis: The Purpose of exploratory analysis is to discover patterns or correlation between explanatory variables. The correlation matrix for the variables in the dataset is calculated. The pair-wise correlation among variables indicated low correlation among most of the variables except for the pairs: 1) “user_listed_counts-user_followers_count” and 2)“geo_coordinates–num_ hashtags”) as shown in Fig. 1(a) and correlation matrix in Fig. 1(b).

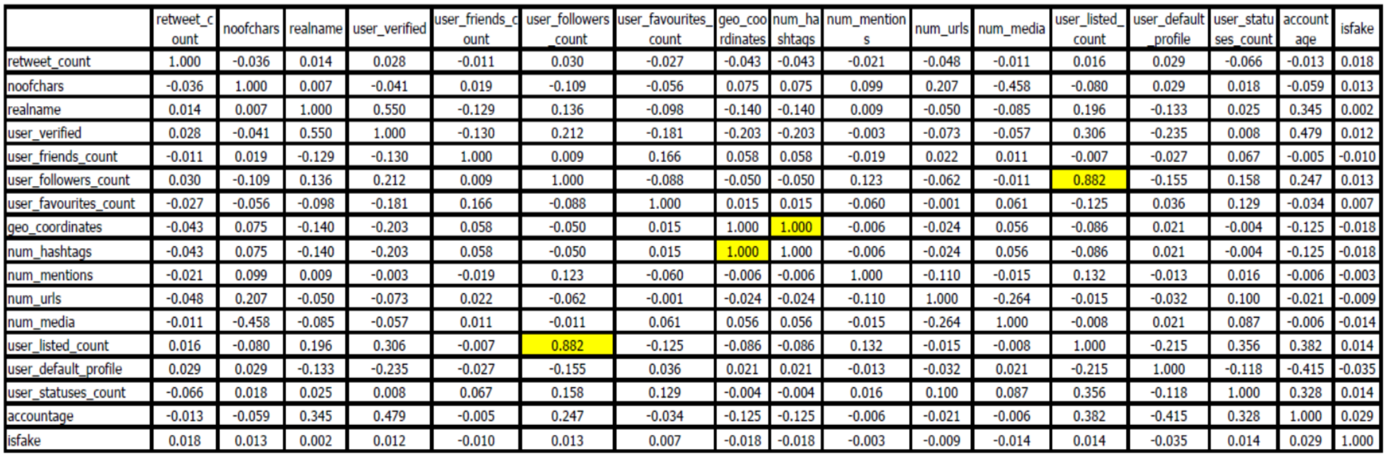
Strong correlation between explanatory variables (collinearity) can result in limitations of the analytical models. Variance inflation factors (VIF) test [25] was applied on data to verify collinearity among explanatory variables. VIF measures **the variance** between two variables **when correlated compared to variance when they are uncorrelated.** VIF value can indicate the degree of collinearity, where, VIF = 1 means variables are not correlated, 1 < VIF < 5 means moderately correlated and VIF >=5 indicates highly correlated variables. Results of VIF test indicated high VIF value for the variables user_followers_count =5.5, user_listed_count=6.77 and “infinity” for the variables (num_hashtags and geo_coordinates) as shown in Table III(a).

TABLE. II. DATASET METADATA

Variable	Type	level	
			Description
tweet_id	Integer	Content	Id for each Tweet.
created_at	Date/Time	Content	Date at which tweets had been created
retweet_count	Integer	Context	Number of time news had been retweeted
Text	Text	Content	The textual content of the news tweet
num_hashtags	Integer	Content	Number of hashtags included in the tweet.
num_mentions	Integer	Content	Number of users who are mentioned in the tweet.
num_urls	Integer	Content	Number of URLs included in the tweet.
num_media	Integer	Content	Number of images/videos included in the tweet.
user_screen_name	Text	User	Account display name
user_verified	Boolean	User	Whether or not the Twitter account is verified.
user_friends_count	Integer	User	The number of friends of the author
user_followers_count	Integer	User	The number of followers the user has.
user_favourites_count	Integer	User	The number of tweets the user has favorited.
tweet_source	Text	User	URL of the tweet
geo_coordinates	Integer	User	The geographic location of the Tweet as reported by the user or client application.
user_default_profile_image	Boolean	User	Whether or not the user uses the default profile image or his account
user_description	Text	User	Description included in the profile
user_listed_count	Integer	User	The number of public lists that the user is a member of.
user_name	Text	User	User’s unique name.
user_profile_use_background_image	Boolean	User	Does the profile has a background image
user_default_profile	Boolean	User	Is this the default user account??
user_statuses_count	Integer	User	Number of tweets issued by the user
isfake	Boolean	Content	Label each tweet as fake/real.



(a) Pairwise Correlation Matrix.



	retweet_count	noofchars	realname	user_verified	user_friends_count	user_followers_count	user_favourites_count	geo_coordinates	num_hashtags	num_mentions	num_urls	num_media	user_listed_count	user_default_profile	user_statuses_count	accountage	isfake
retweet_count	1.000	-0.036	0.014	0.028	-0.011	0.030	-0.027	-0.043	-0.043	-0.021	-0.048	-0.011	0.016	0.029	-0.066	-0.013	0.018
noofchars	-0.036	1.000	0.007	-0.041	0.019	-0.109	-0.056	0.075	0.075	0.099	0.207	-0.458	-0.080	0.029	0.018	-0.059	0.013
realname	0.014	0.007	1.000	0.550	-0.129	0.136	-0.098	-0.140	-0.140	0.009	-0.050	-0.085	0.196	-0.133	0.025	0.345	0.002
user_verified	0.028	-0.041	0.550	1.000	-0.130	0.212	-0.181	-0.203	-0.203	-0.003	-0.073	-0.057	0.306	-0.235	0.008	0.479	0.012
user_friends_count	-0.011	0.019	-0.129	-0.130	1.000	0.009	0.166	0.058	0.058	-0.019	0.022	0.011	-0.007	-0.027	0.067	-0.005	-0.010
user_followers_count	0.030	-0.109	0.136	0.212	0.009	1.000	-0.088	-0.050	-0.050	0.123	-0.062	-0.011	0.882	-0.155	0.158	0.247	0.013
user_favourites_count	-0.027	-0.056	-0.098	-0.181	0.166	-0.088	1.000	0.015	0.015	-0.060	-0.001	0.061	-0.125	0.036	0.129	-0.034	0.007
geo_coordinates	-0.043	0.075	-0.140	-0.203	0.058	-0.050	0.015	1.000	1.000	-0.006	-0.024	0.056	-0.086	0.021	-0.004	-0.125	-0.018
num_hashtags	-0.043	0.075	-0.140	-0.203	0.058	-0.050	0.015	1.000	1.000	-0.006	-0.024	0.056	-0.086	0.021	-0.004	-0.125	-0.018
num_mentions	-0.021	0.099	0.009	-0.003	-0.019	0.123	-0.060	-0.006	-0.006	1.000	-0.110	-0.015	0.132	-0.013	0.016	-0.006	-0.003
num_urls	-0.048	0.207	-0.050	-0.073	0.022	-0.062	-0.001	-0.024	-0.024	-0.110	1.000	-0.264	-0.015	-0.032	0.100	-0.021	-0.009
num_media	-0.011	-0.458	-0.085	-0.057	0.011	-0.011	0.061	0.056	0.056	-0.015	-0.264	1.000	-0.008	0.021	0.087	-0.006	-0.014
user_listed_count	0.016	-0.080	0.196	0.306	-0.007	0.882	-0.125	-0.086	-0.086	0.132	-0.015	-0.008	1.000	-0.215	0.356	0.382	0.014
user_default_profile	0.029	0.029	-0.133	-0.235	-0.027	-0.155	0.036	0.021	0.021	-0.013	-0.032	0.021	-0.215	1.000	-0.118	-0.415	-0.035
user_statuses_count	-0.066	0.018	0.025	0.008	0.067	0.158	0.129	-0.004	-0.004	0.016	0.100	0.087	0.356	-0.118	1.000	0.328	0.014
accountage	-0.013	-0.059	0.345	0.479	-0.005	0.247	-0.034	-0.125	-0.125	-0.006	-0.021	-0.006	0.382	-0.415	0.328	1.000	0.029
isfake	0.018	0.013	0.002	0.012	-0.010	0.013	0.007	-0.018	-0.018	-0.003	-0.009	-0.014	0.014	-0.035	0.014	0.029	1.000

Fig. 1. (b) Correlation Matrix between Dataset Explanatory Variables.

The variable with the highest VIF value is removed from the dataset and the VIF test is repeated as values may change after each variable is removed. Results after removing “geo_coordinates” variable and repeating the test for the 2nd time indicated low VIF for “num_hashatgs” while both “user_followers_count” and “user_listed_count” still have high VIF values as shown in Table III(b).

The variable with the highest VIF value “user_listed_count” was removed and the test was repeated. Results of the 3rd test indicated low VIF value for all the variables as shown in Table III(c).

3) Variable selection: An important step before model training is to select the features with the highest predictive power. For this study, features are evaluated and ranked based on the model in [27]. The model measures the effect of each variable on the target via an iterative variables' permutations process. The model calculates the mean decrease importance of each variable based on which variable is confirmed or rejected. Results of the feature selection model confirmed all the selected variables as shown in Fig. 2 and Table IV.

TABLE. III. (A) VIF VALUES FOR THE 1ST TEST

Variables	VIF
retweet_count	1.013285
noofchars	1.366300
realname	1.467026
user_verified	1.833684
user_friends_count	1.058622
user_followers_count	5.362508
user_favourites_count	1.143870
geo_coordinates	Inf
num_hashtags	Inf
num_mentions	1.060546
num_urls	1.144046
num_media	1.375882
user_listed_count	6.621056
user_default_profile	1.239349
user_statuses_count	1.549332
Accountage	1.776116
Isfake	1.004115

(B) VIF VALUES FOR THE 2ND TEST

Variables	**VIF**
retweet_count	1.014
Noofchars	1.360
Realname	1.456
user_verified	1.852
user_friends_count	1.056
user_followers_count	**5.435**
user_favourites_count	1.109
num_hashtags	1.065
num_mentions	1.055
num_urls	1.148
num_media	1.382
user_listed_count	**6.689**
user_default_profile	1.228
user_statuses_count	1.564
Accountage	1.774
Isfake	1.003

(C) VIF VALUES FOR THE 3RD TEST

Variables	**VIF**
retweet_count	1.013
Noofchars	1.364
Realname	1.474
user_verified	1.815
user_friends_count	1.051
user_followers_count	1.142
user_favourites_count	1.088
num_hashtags	1.062
num_mentions	1.055
num_urls	1.149
num_media	1.380
user_default_profile	1.227
user_statuses_count	1.233
Accountage	1.768
Isfake	1.005

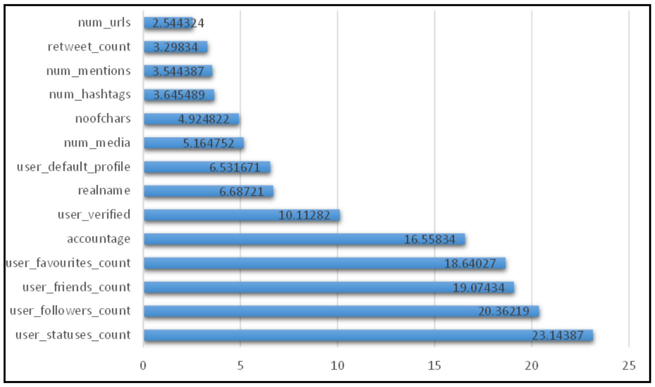


Fig. 2. Mean Importance of the Explanatory Variables.

TABLE. IV. MEAN IMPORTANCE OF THE EXPLANATORY VARIABLES

Variable	**meanImp**	**decision**
user_statuses_count	23.14387	Confirmed
user_followers_count	20.36219	Confirmed
user_friends_count	19.07434	Confirmed
user_favourites_count	18.64027	Confirmed
accountage	16.55834	Confirmed
user_verified	10.11282	Confirmed
realname	6.68721	Confirmed
user_default_profile	6.531671	Confirmed
num_media	5.164752	Confirmed
noofchars	4.924822	Confirmed
num_hashtags	3.645489	Confirmed
num_mentions	3.544387	Confirmed
retweet_count	3.29834	Confirmed
num_urls	2.544324	Confirmed

C. Analytical Models

A set of the most known and most widely used models for fake news detection in the literature were selected for this study. The selected models cover different learning models (linear, non-linear, tree-based and ensemble).

1) Linea- learning models

- LDA: LDA is a linear learning model that tries to find for a grouping of predictors that can discriminate two targets. LDA is related to regression as they both try to express the relationship between one dependent response variable and a set of independent variables. However, LDA uses continuous independent variables and a categorical dependent variable. The label for the new instance is estimated by the probability that inputs belong to each class and the instance is assigned the class with the highest probability calculated based on Bayes Theorem [28].

2) Non-linear learning models

- Mixture Discriminant Analysis (MDA): MDA is an extension of LDA that models the within-group multivariate density of predictors through a mixture (i.e., a weighted sum) of multivariate normal distributions [29]. In principle, this approach is useful for modeling multivariate non-normality or nonlinear relationships among variables within each group, allowing for more accurate classification. to determine whether underlying subclasses may be present in each group.

- SVM: A supervised learning model that analyses data in order to identify patterns. Given a set of labeled training data, SVM represents instances in the dataset as points in a high-dimensional space and tries to identify the best separating hyperplanes between different classes. New instances are represented in the same space and are classified to a specific class based on their closeness to the separating gap [30].
- NB: Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem [31]. It assumes complete variables independence, as the presence/absence of one variable is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB bases its predictions for new observations based on the analysis of previous observations. NB model usually outputs a probability score and class membership.
- KNN: KNN is an Instance-based or memory-based learning, labeling new instances is based on in-memory instances stored in advance. In KNN, no internal model is constructed, and computations are performed at classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. The instance is labeled with the class most common among its neighbors. KNN determines neighbors based on distance using Euclidian, Manhattan or Murkowski distances for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training instances (k) that are the closest to the new point and label is assigned based on them [32].

3) ANNs: ANNs try to mimic the performance of the biological neural network of the human brain. ANNs are adaptive, fault tolerant and can learn by example. An ANN is composed of a set of connected neurons organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs. The weights on the connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANNs come with many structures. The most common structures are feed-forward neural network (single and multi-layer) and recurrent neural nets. Multilayer perceptron (MLP) is a feed-forward ANN that contains at least one hidden layer. Neurons in each layer use supervised learning techniques [33]. LVQ is also a feed-forward ANN that is based on the winner – takes – all learning approach. In this approach, the distance is measured between each data point and the output. The smaller distance indicates a winner which is then adopted by adjusting its weights. It's as if, the prototype is moved closer if it correctly classifies the data point or moved away if otherwise [34].

4) Tree-based learning: Tree-based learning makes use of decision trees as a predictive model. Items are represented in a tree structure. In such structure, nodes represent test points for variables, leaves represent class labels and branches represent a combination of variables that lead to class labels [35]. Two popular implementations of DTs are a) CART [36] and C50 [37]. CART is a binary DT that can be used for classification and regression. For classification, CART used Gini index function to indicate the purity of the leaf nodes. C5.0 algorithm is used to build decision tree or a rule set. It works by splitting the sample based on the field that provides the maximum information gain. It uses subsamples based on a variable and iteratively split data until subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed/pruned.

5) Ensemble learning: Ensemble learning trains multiple models using the same learning algorithm and set learners to solve the problem. The main causes of error in learning are due to noise, bias, and variance. Ensemble minimizes these factors and may produce a more reliable classification than a single classifier. Bagging (i.e. Bagging CART, Random Forest) and Boosting (i.e. Ada Boost and Stochastic Gradient Boosting) get N learners by generating additional data in the training stage. N new training data sets are produced by random sampling with replacement from the original set. By sampling with replacement, some observations may be repeated in each new training data set. In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often. Both are good at reducing variance but only Boosting tries to reduce bias, Bagging may solve the over-fitting problem, while Boosting can increase it [38].

D. Performance Evaluation Metrics

The performance of the selected models' predictive power is evaluated based on accuracy, precision, recall, and F-measure (F1).

1) Accuracy: Indicates the ability of the model to differentiate the fake and real instances correctly. It’s the proportion of true positive (TP) and true negative (TN) in all evaluated news:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

where

TP: is the total number of tweets correctly identified as fake.

FP: is the total number of tweets incorrectly identified as fake.

TN: is the total number of tweets correctly identified as real.

FN: is the total number of tweets incorrectly identified as real.

2) Precision and recall: Precision and recall can give a better insight into the performance as they do not assume equal misclassification costs. Precision indicates is the fraction of tweets correctly classified as fake among all classified instances, while recall is the fraction of tweets correctly classified as fake over the total number of fake tweets. relevant instances.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

3) F-measure: F-measure (F1) is calculated based on a combination of both precision and recall to provide a better evaluation of predictive performance.

$$F_1 = \frac{2\,x\,Precision\,x\,Recall}{Precision+Recall} \qquad (4)$$

IV. Model Training and validation

Model training is an important step, as based on which models will behave. During this step, models are fed with labeled training dataset. Dataset was split into 80% for training and 20% for testing. For model training, 5 x 2-fold cross-validation was applied as recommended by [26]. Initial parameters are tuned via grid search during the training stage. The optimal parameter values are selected based on cross-validated accuracy as shown in Table V and the mean accuracy achieved by the models during the cross-validation is shown in Fig. 3.

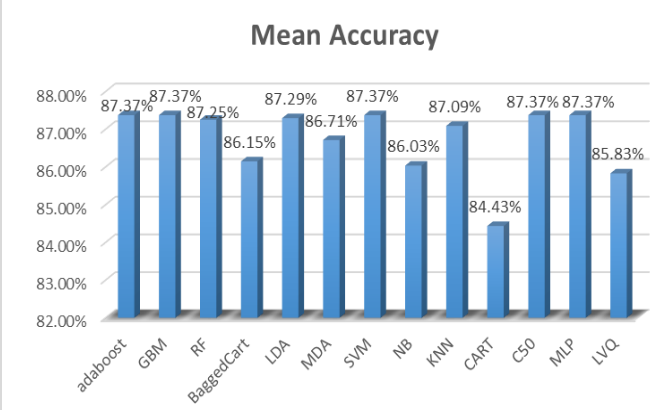


Fig. 3. Mean Accuracy Values During Cross Validation.

TABLE. V. Parameters Values

Model	Parameters	Tuning values
AdaBoost	Number of iterations	[**50**,100,150]
	Maximum depth	[**1**,2,3]
	Weight update coefficient	**Beriman**
C50	Trials(number of iterations)	**1**
	Model(tree-based or rule-based)	**Rules**
	Winnow(use feature selection?)	**True**
DT	Complexity parameter	[0.0005500550, 0.0007616146, **0.0012376238**]
GBM	Number of trees	[**50**,100,150]
	Interaction depth(number of splits)	[**1** , 2 , 3]
	Shrinkage(learning rate)	0.1
	Min observations in node	10
KNN	K	[5 , 7 , **9**]
LVQ	Learning capacity(size)	[6, **9** , 12]
	K	[1 , 6 , **11**]
MDA	Subclasses(#*Subclasses* Per Class)	[**2**, 3 , 4]
MLP	Learning function	Std_Backpropagation
	Maximum iterations(maxit)	100
	Initial weight matrix (initFunc)	Randomized_Weights
	number of units in the hidden layer(size)	[**1**, 3 , 5]
SVM	δ	0.08984069
	C (cost of penalty)	[**0.25**, 0.50 , 1.00]

V. Results and Discussion

The experiment was carried out on an Acer machine with 64-bit Windows 10 OS, Intel® Core™ i7 – 7500U CPU @ 2.70GHZ and 8 GB Memory using R language. In order to test the performance of the selected models, unlabeled 20% of the dataset was used as an input the trained models for performance evaluation. Results of testing are used to compare the models based on a) predictive performance in terms of the selected metrics, and b) amount of time and memory usage during processing and classification time.

A. Predictive Power Evaluation

Results in Table VI show that linear-based learning model LDA achieved high performance compared to other models with 86.41% accuracy, 86.41% precision, 100% recall and 92.71% F1. Within the non-linear classifiers, SVM outperformed other non-linear models with 86.41 % accuracy, followed by MDA with 86.07%, KNN with 85.82% where NB achieved the lowest accuracy of 85.57%. SVM also outperformed the other non-linear models in both recall and F1 values followed by MDA, KNN and finally NB. However, KNN outperformed all non-linear models with precision of 86.45% followed by SVM with 86.41%, MDA with 86.36% and finally NB with 86.35%. It worth noting that NB model works well only with categorical data and cannot perform on continuous data. Thus, discretizing the continuous data may lead to better performance of this model.

For ANNs – despite achieving 86.41% accuracy during training, LVQ accuracy dropped to 82.49% to achieve approximately 4% lower accuracy, 6% lower recall and 2% lower F1 compared to MLP where precision of the two models is almost the same with 86.41% for MLP and 86.36% for LVQ. For tree-based learning models both CART and C50 trees achieved the same performance over all metrics with 86.41% accuracy, 86.41% precision, 100% recall and 0.9261 F1. For ensemble learning – based models boosted models (GBM and AdaBoost) showed higher accuracy, recall, and F1 compared to bagged models (BaggedCart and RF) with 86.41% accuracy, 100% recall, and 92.71% F1. However, BaggedCart achieved 86.59% precision which outperforms all the ensemble-learning models. Comparison between different models is shown in Fig. 4.

B. *Operational Characteristics Evaluation.*

Beside their predictive capabilities, operational characteristics in terms of runtime and memory usage were tested for each model during both processing and classification as shown in Table VII. the running statuses of each model was obtained using “profvis” profiling tool in R. Results show variation in time and memory consumption as Adaboost has the maximum processing time which is much longer than all other models recording *1 hour 48 seconds and 350 milliseconds* while, the processing time of all other models ranged from 350 milliseconds for LDA (lowest processing time) to 42 seconds, 450 milliseconds for RF. For non-linear models, KNN achieved the lowest processing time during training, followed by MDA, SVM, and finally NB while MDA achieved the lowest classification time followed by KNN, SVM, and NB. For memory usage, KNN had the minimum usage during training and classification followed by MDA, NB, and finally SVM. For tree-based models, C50 outperformed CART in training time while they both achieved the same classification time. For memory usage, CART had the lowest memory usage. For ANNs, LVQ outperformed MLP with lower time and memory usage in both phases. For ensemble learning models BaggedCart achieved the lowest processing time while GBM achieved the lowest classification time and memory usage during both training and classification among the rest of the models. It is worth noting that despite their high processing time, AdaBoost achieved reasonable classification time in relevance with the ensemble learning models. The best classification time was achieved by GBM and LVQ (10 milliseconds), followed by CART, C50 and MDA (30 milliseconds). LDA had the lowest memory usage during classification, followed by KNN they both had less than 200 MB memory usage. In runtime, MLP and LVQ achieved the lowest memory usage followed by KNN and LDA.

A comparison between the models based on time and memory usage is found in Fig. 5(a,b,c,d).

Choosing the suitable model has to balance between high predictive performances, low classification time and memory usage. That’s why LDA and CART can be recommended as they provide high predictive power with low time and memory usage compared to other models. GBM is recommended too as it gives a good balance with the same performance with lower classification time and memory but higher processing time and memory.

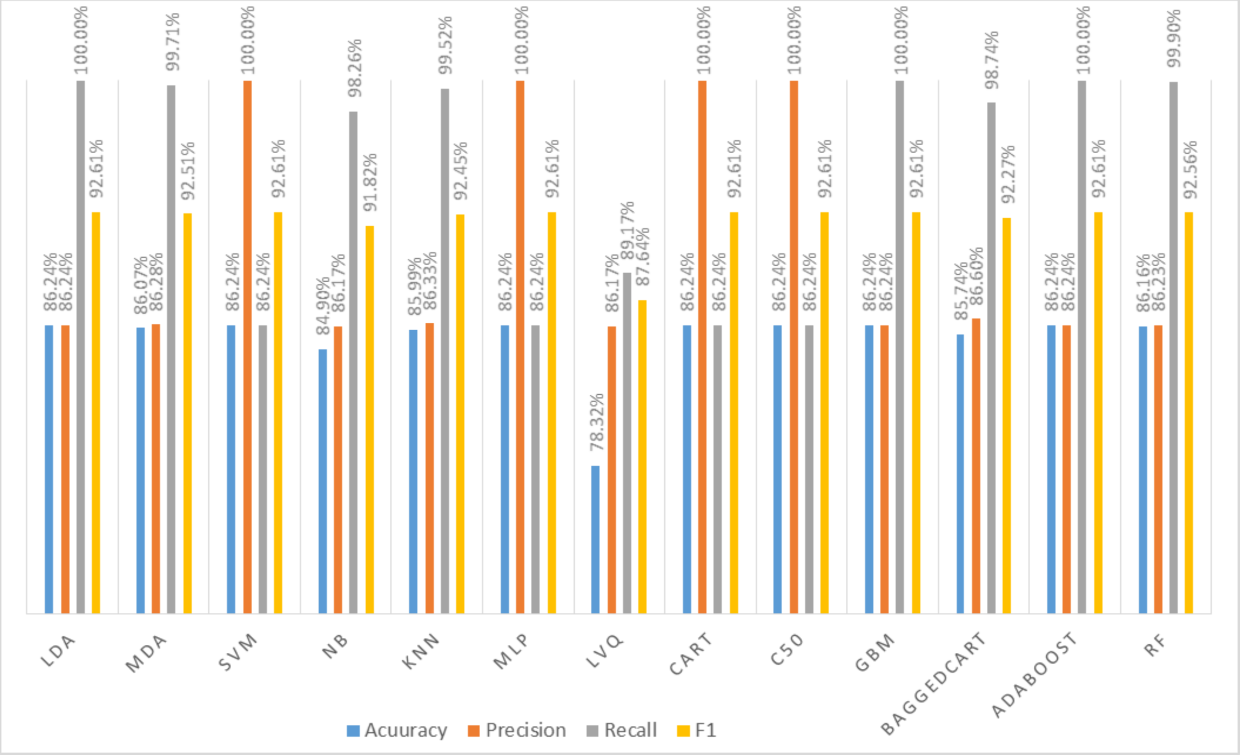


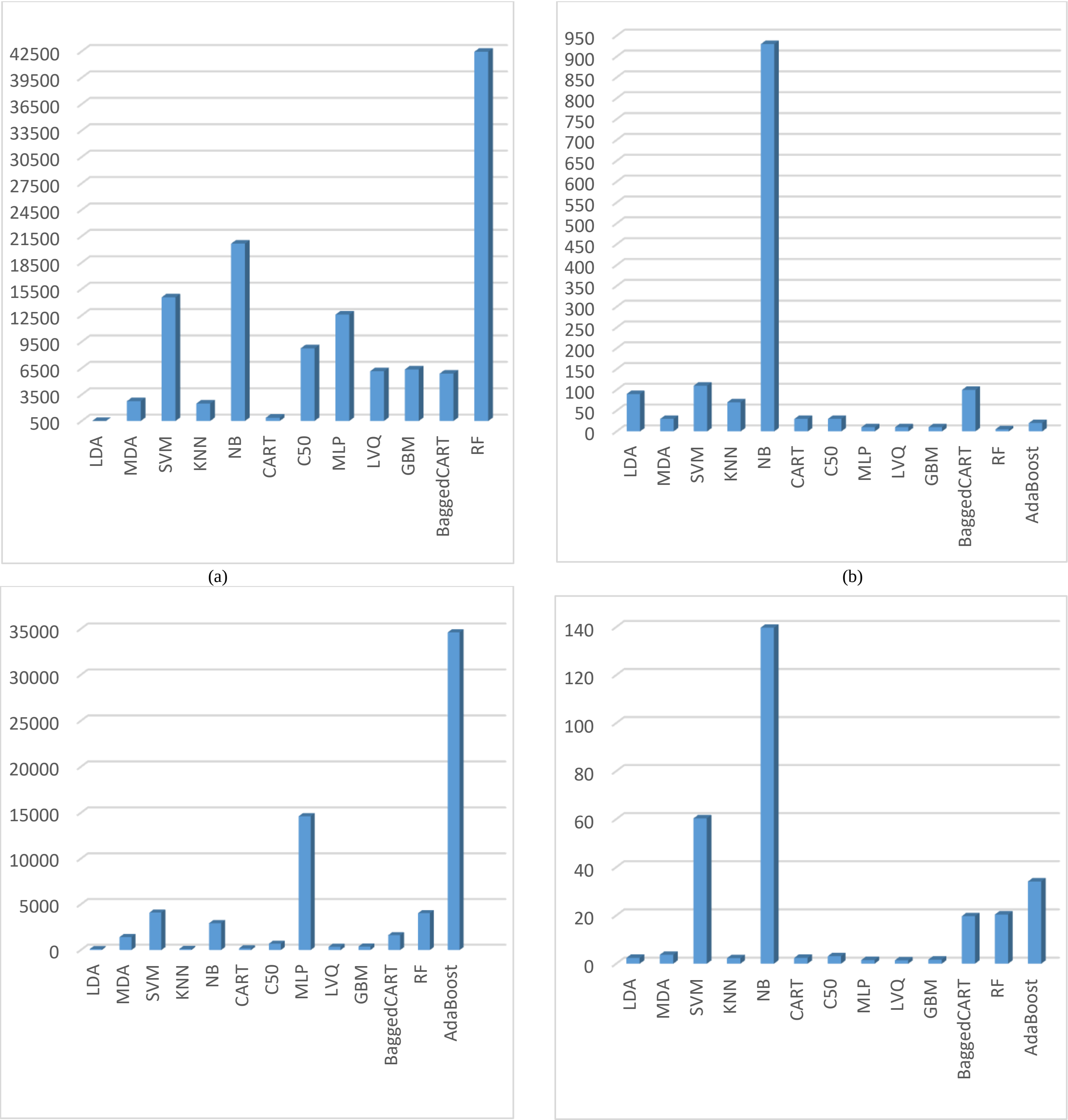
Fig. 4. Performance Comparisons of the Models.

TABLE. VI. MODELS' PERFORMANCE

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Linear Based learning				
LDA	**0.8641**	0.8641	1	0.9271
Non-Linear Learning				
MDA	0.8607	0.8636	0.9961	0.9251
SVM	**0.8641**	0.8641	**1**	**0.9271**
NB	0.8557	0.8635	0.9894	0.9222
KNN	0.8582	**0.8645**	0.9913	0.9236
ANN				
MLP	0.8641	0.8641	1	0.9271
LVQ	0.8249	0.8636	0.9469	0.9003
Tree-based Learning				
CART	0.8641	0.8641	1	0.9271
C50	0.8641	0.8641	1	0.9271
Ensemble learning				
GBM	**0.8641**	0.8641	**1**	**0.9271**
BAGGEDCART	0.8549	**0.8659**	0.9846	0.9214
AdaBoost	**0.8641**	0.8624	**1**	**0.9271**
RF	0.8632	0.8645	0.9981	0.9265

TABLE. VII. TIME AND MEMORY CONSUMPTION OF THE MODELS

Model	**Processing**		**Classification**	
	Time(ms)	**Memory(MB)**	**Time(ms)**	**Memory(MB)**
Linear Learning				
LDA	**00:00:00.350**	**104**	**00:00:00.090**	2.4
Non-linear Learning				
MDA	00:00:02.750	1419.4	**00:00:00.030**	3.7
SVM	00:00:14.550	4097.7	00:00:00.110	60.3
NB	00:00:20.650	2941.5	00:00:00.930	139.7
KNN	**00:00:02.480**	**119.6**	00:00:00.070	**2.3**
Tree-based Learning				
CART	00:00:00.870	**180**	**00:00:00.030**	2.4
C50	**00:00:08.750**	690	00:00:00.030	3.1
ANN				
MLP	00:00:12.580	14629.4	00:00:00.010	1.5
LVQ	**00:00:06.150**	**355.3**	**00:00:00.010**	**1.4**
Ensemble Learning				
GBM	00:00:06.360	**380.4**	**00:00:00.010**	**1.7**
BaggedCART	**00:00:05.880**	1624.6	00:00:00.100	19.7
AdaBoost	01:00:48.530	34620.7	00:00:00.110	4.1
RF	00:00:42.450	4032.2	00:00:00.050	20.4



(a)

(b)

(c)

(d)

Fig. 5. (a) Processing Time of Models[3]. (b) Classification Time of Models. (c) Processing Memory usage. (d) Memory usage During Classification.

[3] AdaBoost Processing time is not included due to its large value compared to other models (3648.53 second).

VI. Conclusion and Future Work

This study tries to present an evaluation of the performances of different data mining models for credibility assessment in the context of social media. This study focused on Twitter news credibility assessment as a case study. The bulk of works in the literature focused on identifying the most informative features, feed those features into different models to select the model with higher predictive power and all of them disregarded time and memory consumption during both processing and runtime. Results of these studies contrast each other and cannot give a unified decision. This study tries to address this limitation by benchmarking different data mining models for news credibility assessment on Twitter. Models are evaluated in terms of their predictive performance using Accuracy, Precision, Recall and F-measure and time and memory usage during both processing and prediction.

However, the study still has some limitations and future research opportunities. First, the results on Twitter data may not be applicable on different social media contexts (i.e. blogs, Facebook, etc.). One possible future research shall utilize different datasets in different contexts for the evaluation. Another possible future work can be to explore the performance of other models including the less well known models and deep learning models. Performance can be evaluated with missing and noisy labels.

References

[1] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, Karl Aberer, “Web Credibility: Features Exploration and Credibility Prediction”, in the proceedings of European Conference on Information Retrieval. ECIR 2013:Advances in Information Retrieval pp 557-568, 2013.

[2] John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, Sibel Adal, “Credibility in Context: An Analysis of Feature Distributions in Twitter”, In the proceedings of the International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 2012.

[3] A. A. A. Mansour, "Labeling Agreement Level and Classification Accuracy," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 271-274. doi: 10.1109/SITIS.2016.51

[4] Dana Movshovitz-Attias, Yair Movshovitz-Attias,Peter Steenkiste, Christos Faloutsos, “Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow”. In the Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining Pages 886-893. 2013.

[5] Ruohan Li, Ayoung Suh , “Factors Influencing Information credibility on Social Media Platforms: Evidence from Facebook Pages”, In the proceedings of the 3rd Information Systems International Conference (ISICO2015), 2015.

[6] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter”, In the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011.

[7] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, Julia Schwarz, “Tweeting is Believing? Understanding Microblog Credibility Perceptions”, CSCW 2012, USA.

[8] Kanda Runapongsa Saikaew, Chaluemwut Noyunsan, “Features for Measuring Credibility on Facebook. Information”. In the proceedings of the XIII International Conference on Computer Science and Information Technology (ICCSIT 2015),Thailand, 2015.

[9] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro, “Some Like it Hoax: Automated Fake News Detection in Social Networks”, CoRR,abs/1704.07506, 2017.

[10] Mehrbod Sharifi, Eugene Fink, and Jaime G. Carbonell. “Detection of Internet scam using logistic regression”. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pages 2168–2172, 2011.

[11] James Fairbanks, Natalie Fitch, Nathan Knauf, Erica Briscoe, “Credibility Assessment in the News: Do we need to read?”, MIS2’18, Feb 2018, Los Angeles, California USA.

[12] William Ferreira and Andreas Vlachos. “Emergent: a novel dataset for stance classification”. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[13] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media", in Proceedings of International Conference on Data Mining, pp. 103-1108, 2013.

[14] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. “Enquiring minds: Early detection of rumors in social media from enquiry posts”. In Proceedings of the 24th International Conference on World Wide Web . ACM, 1395–1405, 2015.

[15] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. “Detect rumors using time series of social context information on microblogging websites”. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 1751–1754. 2015.

[16] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy”. In Proceedings of the 22nd international conference on World Wide Web. ACM, 729–736. 2013.

[17] Manish Gupta, Peixiang Zhao, Jiawei Han, “Evaluating Event Credibility on Twitter”, Proceedings of the 2012 SIAM International Conference on Data Mining, pages = 153-164.

[18] Rim El Ballouli, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem Hajj and Khaled Shaban, “CAT: Credibility Analysis of Arabic Content on Twitter”, Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 62–71, 2017.

[19] Sahar. F. Sabbeh, S. Batawah, Arabic news credibility on Twitter: An Enhanced Model using Hybrid Features", Journal Of Theoretical And Applied Information Technology , Vol 96 April 2018.

[20] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, Fabiana Zollo, “Polarization And Fake News: Early Warning Of Potential Misinformation Targets”, Arxiv:1802.01400v1 [Cs.Si] 5 Feb 2018.

[21] R.Deepa Lakshmi , N.Radha , “Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools “, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 09, p= 2783-2789, 2010.

[22] Marin Vuković ,Krešimir Pripužić, Hrvoje Belani, “An Intelligent Automatic Hoax Detection System”, In Knowledge-Based and Intelligent Information and Engineering Systems , pages 318–325. Springer, Berlin, Heidelberg, September 2009.

[23] Benjamin Markines, Ciro CaŠuto, and Filippo Menczer, “Social spam detection. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web . ACM, 41–48, 2009.

[24] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, Benno Stein” A Stylometric Inquiry into Hyperpartisan and Fake News”, arXiv:1702.05638, 2017.

[25] James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert, “ An Introduction to Statistical Learning (8th ed.)”. Springer Science+Business Media New York. ISBN 978-1-4614-7138-7, 2017.

[26] Dietterich, T. G. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”. Neural Comput, 10(7):1895–1923, 1998.

[27] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki, "Boruta – A System for Feature Selection", Fundamental Informaticae volume101, pages:271–285, 2010.

[28] McLachlan, G. J. Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469. 2004.

[29] Fraley, C., & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611-631.2002.

[30] Cortes, Corinna; Vapnik, Vladimir N. "Support-vector networks". Machine Learning. Volume: 20 No:3: p:273–297. doi:10.1007/BF00994018, 1995.

[31] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955. 2003.

[32] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

[33] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors);

[34] Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.

[35] T. Kohonen, "Learning vector quantization", in M.A. Arbib, The Handbook of Brain Theory and Neural Networks, Cambridge, MA: MIT Press, pp. 537–540, 1995.

[36] Rokach, Lior; Maimon, O. “Data mining with decision trees: theory and applications”. World Scientific Pub Co Inc. ISBN 978-9812771711. 2008.

[37] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J., “Classification and regression trees”. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8. 1984.

[38] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1-2): 1–39.

AUTHOR’S PROFILE

SAHAR F. SABBEH earned her B.Sc. degree in information systems from the Faculty of computers and information technology, Mansoura university, Egypt in 2003. She earned her M.Sc. also in information systems from the same department in 2008 and earned her Ph.D. degree in 2011 She has been a member of the IEEE since 2017. Dr. Sabbeh worked in Alzarka High Institution for management information systems from 2004 to 2009. She worked at Misr Higher Institution of Engineering and Technology, Mansoura, Egypt from 2009 till 2011. She worked with the Faculty of Computers and Information Technology, Banha University, Egypt during the period from 2011 - 2018 as an assistant professor. She also worked part time as an assistant professor in several reputable private universities in Cairo, Egypt. She worked as an Associate professor in the faculty of computers and information technology, King Abdul-Aziz university, KSA during the period from 2016–2018. Currently, she is an associate professor at the Faculty of Computers and Information Technology, Banha University, Egypt and an associate professor in the computer science and engineering, university of Jeddah, KSA. She supervised 5 M.Sc. and one Ph.D. students.