# Improving Credit Scorecard Modeling Through Applying Text Analysis

Omar Ghailan
Faculty of Computers & Info.
Cairo University, Egypt

Hoda M.O. Mokhtar
Faculty of Computers & Info.
Cairo University, Egypt

Osman Hegazy
Faculty of Computers & Info.
Cairo University, Egypt

***Abstract*—In the credit card scoring and loans management, the prediction of the applicant's future behavior is an important decision support tool and a key factor in reducing the risk of Loan Default. A lot of data mining and classification approaches have been developed for the credit scoring purpose. For the best of our knowledge, building a credit scorecard by analyzing the textual data in the application form has not been explored so far. This paper proposes a comprehensive credit scorecard model technique that improves credit scorecard modeling though employing textual data analysis. This study uses a sample of loan application forms of a financial institution providing loan services in Yemen, which represents a real-world situation of the credit scoring and loan management. The sample contains a set of Arabic textual data attributes defining the applicants. The credit scoring model based on the text mining pre-processing and logistic regression techniques is proposed and evaluated through a comparison with a group of credit scorecard modeling techniques that use only the numeric attributes in the application form. The results show that adding the textual attributes analysis achieves higher classification effectiveness and outperforms the other traditional numerical data analysis techniques.**

***Keywords*—*Credit Scoring; Textual Data Analysis; Logistic Regression; Loan Default.***

## I. Introduction

Credit Scoring is a decision support tool used to identify the level of risk associated with the applicants for a specified service. It is based on applying a group of statistical techniques to predict the behaviour of those applicants and assigning scores reflecting how much good or bad they are expected to be [1]. The credit scorecard models are widely used in the risk management of the banks, insurance companies, and other financial institutions aim to identify the quality or the risk of their customers. The credit scorecard model is designed to replace the old judgmental system which depends on the decision maker or the creditor to assign the risk score. The credit scoring model's purpose is to increase the efficiency, and the reliability of the judgment process [2].

The developed applications and the proposed researches in this area used several statistical techniques to build the credit scorecard models such as Support Vector Machine [3][4], Neural Networks [5][6], Logistic Regression [7][8], Genetic Programming [9], Nearest Neighbour [10], and other hybrid techniques [11][12]. Each of those techniques has its form in representing the scorecard generated from the model. Each technique has its strengths and its advantages in some of the circumstances but there is no overall best one in all the circumstances. Group of the existing credit scoring models will be discussed in section II.

The data sources used in making a credit scoring decisions includes both the applicant form details and the information collected by a credit reference agency like the public registries, internal records in the bank from previous experiences, transactions, and any other activities initiated by the applicant in the bank [13].

Beside all the work that has been done in this area, there is a drawback point related to the structure of the data used to build the statistical model. This limitation was a result of using only that data that could be represented as numeric values and neglecting the textual data regardless of its importance in purpose of simplifying the analysis calculations and the shape of the output of the model. This limitation results in decreasing the degree of efficiency of those models by neglecting some of the available valuable data that could be used to extract some features to increase the percentage of accuracy and correctness of the classification model.

In this study, we aim to improve the existing credit scorecard modeling by employing the textual data analysis. The textual analysis results in extracting a group of textual features inserted to the credit scorecard to increase its accuracy depending on statistical aspects. The resulting scorecard from the proposed methodology is compared with a group of scorecards generated from using only the numerical data analysis. Logistic Regression, Decision Tree, SVM, and Neural Network techniques are used in the comparison using a sample of 180 loan application forms collected from a financial institution providing loan services in Yemen. The results of the comparison using Error Rate, Recall, Precision, and F1-score show the improvement in the credit scorecard accuracy when employing the textual data analysis.

The organization of this paper is as follows, related works are discussed in Section II. The proposed methodology of integrating the textual data analysis in the credit scorecard model is discussed in Section III. In Section IV, experimental results are explained. Finally, section V draws the conclusions of the paper and highlights future work related to the study.

## II. Related Works

### A. Related Work in Credit Scoring

Recently, the credit scoring statistical techniques have been investigated widely due to increasing the interest of financial institutions to classify their customers specially that related to the loans. Several studies have been conducted to improve the accuracy and effectiveness of the classification techniques used in building the credit scoring models.

In [3], the authors proposed a hybrid credit scoring model by integrating SVM technique with Linear Discriminant Analysis (LDA), F-score, Decision Tree, and Rough Sets as features selection pre-processing methods. The experiments using Australian training dataset extracted from UCI Repository concluded that the hybrid model increases the classification accuracy with average accuracy rate approaches 86.52% when applying SVM/LDA model.

Another Hybrid SVM-based credit scoring models was investigated in [4]. The results showed that integrating SVM and genetic algorithm techniques can enhance the feature selection task compared to decision tree and neural networks classifiers.

The authors of [5] investigated building credit scoring models using neural networks classification techniques such as multilayer perceptron and modular neural networks compared to the other traditional techniques such as logistic regression and linear discriminant analysis. The results indicated that customized neural networks model with total correct classification rate approaches 83.19% performs better than the other models have been used in the comparison.

Another study investigated applying neural networks technique in the credit scorecard modeling was represented in [6]. A comparison with other techniques such as Probit Analysis, Discriminant Analysis, and Logistic Regression was conducted to evaluate the NN model's performance using credit risk datasets collected from Egyptian banks. The results concluded that neural nets model outperformed the other techniques with accuracy rate approaches 95.52%.

The authors in [7] proposed two credit scoring models using Logistic Regression and Radial Basis Function techniques applied to training datasets collected from Jordanian banks. The results indicated that the logistic regression model outperformed the radial basis function model with average correct classification accuracy rate approaches 85.4%.

The performance of the Logistic Regression technique when dealing with the credit scoring was investigated in [8]. The authors studied two logistic regression models on a training datasets collected from a Brazilian bank. The study concluded that there is no remarkable improvement in the prediction power when using the logistic regression with state-dependent sample selection model compared to the naive logistic regression model.

Genetic programming (GP) credit scoring model was investigated in [9]. The presented experiments used a collection of Egyptian public sector banks' data sets to test the performance of the proposed model. The experimental results concluded outperforming the GP model compared to the Probit Analysis (PA) Logistic Regression model.

Building a credit scoring model using a hybrid Adaptive Neuro Fuzzy Inference System was proposed in [11]. Using training datasets collected from an international bank in Turkey, the proposed model was compared to the other commonly utilized models in this field. The experimental results concluded that the proposed hybrid model outperforms the other techniques used in the comparison such as Neural Network and Linear Discriminant Analysis models.

The authors in [14] investigated several data mining techniques to study the classification models applied to the imbalanced credit scoring data sets. The authors have explored the suitability of least square, support vector machines, gradient boosting and random forests techniques beside other classification techniques such as logistic regression, neural networks and decision trees. The experiments illustrated that the gradient boosting and random forest classifiers are the most effective techniques for the imbalanced dataset classification.

A reassigning credit scoring model (RCSM) was presented in [15]. The authors constructed a hybrid model using Case-Based Reasoning (CBR) and Artificial Neural Network (ANN) classification techniques. The experimental results concluded outperforming the proposed model with average accuracy rate approaches 82.5% compared to Classification Tree (CART), Linear Discriminant Analysis (LDA), Back Propagation Network (BPN), and Logistic Regression (LR) models.

### *B. Related Work in Arabic Text Categorization*

Due to the increasing in the interests of extracting the information from the textual data to support the decision making process, text mining field expanded lately to increase the efficiency and accuracy of the developed models and to include some new languages that were not targeted before[16].

For Arabic textual data classification, there are many approaches that been investigated towards developing a classifying model depending on traditional classification techniques such as Decision Trees, Logistic Regression, SVM , and neural network techniques [17].

Some researchers developed specially designed models targeted to improve the Arabic text classifiers such as the rule based models generating IF-Then rules based on the Decision Trees Models and this type of models in many studies outperformed the other techniques in case of the Arabic textual data analysis [18].

According to the research in [19], the authors used an Arabic text classifier based on Support Vector Machine technique. The classifier used CHI square method to select the features, which improved the performance of the classifier with F-measure=88.

The authors in [20] compared between Naive Bayesian method and Support Vector Machine algorithm on different Arabic text. The study concluded that the SVM algorithm outperforms the Nave Bayesian model (NB) with regards to all measures used in the comparison.

A comparative study investigated three classifiers for Arabic text categorization in [21]. The results of the comparison showed that the Nave Bayesian model outperforms both the K-NN and the distance-based classifiers.

A distance-based classifier for Arabic text categorization was proposed in [22]. Authors proposed a classifier applied to features extraction for category-specific features that capture inherent category-specific properties. The results showed that the proposed classifier is very accurate and robust with average error rate approaches 0.0744.

Another comparative study in classifying the Arabic text documents using the N-gram frequency statistics was investigated in [23], the authors compared between using the Dice's measure of similarity and Manhattan distance statistics. The study concluded that N-gram text classifier using the Dice measure outperforms the other classifier that used the Manhattan measure.

In [24], a KNN model has been applied to classify Arabic text documents. The authors concluded that using N-Gram in the document indexing outperforms the traditional single term indexing method with average accuracy 0.73 for the N-Gram and 0.66 for the single term indexing technique.

Arabic text classification using Decision Trees (C4.5), One Rule, Rule Induction (RIPPER), and Hybrid (PART) models were studied in a comparative study represented in [25]. The results indicate that PART hybrid approach outperformed the other algorithms used in the study.

In [26], the authors used the decision tree technique based on term stemming, document normalization, and term weighting. Combining Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency Inverse Document Frequency (TFIDF), and pruning infrequent terms significantly affects the classification model by reducing the dimensionality and utilizing the text mining model especially for the large datasets.

A classification system based on Decision Trees algorithm has been evaluated in [27]. The experiments was performed over self-collected datasets and concluded that the proposed hybrid approach using the embedded information gain criterion of the decision tree algorithm is a good Arabic text classifier with average classification accuracy rate approaches 93%.

In [28], a comparative study investigated SVM and Decision Tree C4.5 models using 17658 self-collected documents. The results indicated that Decision Trees Model has more accurate classification than SVM when dealing with the Arabic textual Data.

An association-rules based classifier model for the Arabic textual data has been studied in [29]. The experiments on a self-collected sample concluded that the proposed classifier features high accuracy rates.

A comparative study investigated the performance of Nave Bayes, SVM, and Decision Tree (C4.5) Classifiers when applied to self-collected Arabic text datasets in [30]. The study concluded that the Nave Bayes classifier with average accuracy rate approaches 85.25% outperformed the other models used in the study.

According to [31], a modified Neural Network Model is developed using the Singular value Decomposition representation (SVD). The data used in that NN research consists of 453 documents with 14 categories collected from Al-Hadeeth Books. The modified version of NN in this research outperformed the original artificial network model with classification accuracy rate approaches 88.33%.

## III. METHODOLOGY

This section demonstrates the proposed methodology to improve the credit scorecard model by applying the text data analysis along with traditional numeric data analysis method.

### A. Textual Data Pre-Processing

For the text data, the proposed method of the text analysis consists of the following steps:

- Text Parsing: This step is used to parse, stem, identify the noun groups, and identify the part-of-speech of the text fields. It is also used to remove the words included in the stop list.
- Text Filter: This step is used to filter the words extracted from the text parsing based on some predefined criteria. In our study, the pre-defined minimum number of documents to accept the word is equal 3. Entropy Term Weight formula 1 is used in calculating the importance weight of the extracted words. This technique gives a higher weight for the rare terms. If the term appears in only one document, it will have entropy weight =1 (the max entropy weight). If the term appears in all the documents then it will have entropy weight =0 (the lowest entropy weight) [32]. Entropy Term Weight equation is

$$G_i = 1 + \sum_{j=1}^{d_i} \frac{p_{ij} log_2(p_{ij})}{log_2(n)} \tag{1}$$

  Where: $p_{ij}$= the frequency that term $i$ appears in document $j$ divided by the frequency that term $i$ appears in document collection
  $n$ = number of documents in the collection
  $d_i$= number of documents in which term i appears
- Topics Extraction: A group of topics are constructed from the words resulting from the filtering step. The extracted topics contain the words related to each other based on their appearance in the training dataset collection. The main target of this step is to reduce the dimensionality of the features that will enter the regression model since it will not be effective if we use each extracted word as a feature in our classification model. The employed technique uses Latent Semantic Indexing (LSA) concept through Singular Value Decomposition (SVD) [33]. This helps in grouping similar words into a limited number of distinct sets. Those sets are applied as topics. Each topic will be used as a feature in our regression model. In our research, 30 text topics are collected from each text field to be used in building the credit scorecard.

### B. Interactive Grouping

This step is used to eliminate the weak characteristics that need to be neglected when applying the logistic regression model since some of the used characteristics, either the textual or the numerical fields, have no influence in the final scorecard.

Each numerical field is distributed into intervals based on the similarity in its values' predictive power. Those intervals will be used as features in the final credit scorecard model.

In the developed model, the Information Value is calculated using formula 2 to determine the overall predictive power of the attribute. The predictive power increases as the ability of separating the good and bad records increases [34].

$$IV = \sum_{i=1}^{L}(Distr\ Good_i - Distr\ Bad_i) * ln(\frac{Distr\ Good_i}{Distr\ Bad_i}) \tag{2}$$

Where: $L$=Number of intervals (levels) in the characteristic.

In this study, the characteristics with $IV$ less than 0.10 are eliminated because of their low prediction power.

### C. Applying the Logistic Regression Model

By applying the previous pre-processing steps, each text topic or numerical interval is transformed into a column with value 0 or 1 indicating if the customer's profile contains that feature or not. The resulted 0/1 matrix entered the standard Logistic Regression model represented in 3 which serves the target of generating the credit scorecard for both the numerical attributes and the modified textual attributes [35].

$$Logit(p) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{3}$$

Where: $p$ = the posterior probability of the class 0
$X$ = the input variables
$\beta$ = the estimation coefficients of the X input variables.

### D. Calculating the Final Scores

The Logistic Regression model's output is tuned by applying Weight of Evidence. $WOE$ enhances the final credit scores because Logistic Regression model considers both the features' values 0 and 1 in its processing while the credit scorecard model gives a higher consideration to having the value 1 which indicates that the customer's profile contains the extracted feature.

$$WOE = ln(\frac{Distr\ Good_i}{Distr\ Bad_i}) \tag{4}$$

$$FinalScore = WOE * Estimation\ Coefficient \tag{5}$$

Figure 1 summarizes the proposed methodology.

## IV. EXPERIMENTAL RESULTS

In this section, we study the effect of applying the text analysis on the credit scorecard model's accuracy by comparing the proposed model's results with the traditional techniques that depend only on the numerical variable when building the credit scorecard.

The dataset used in the experiment consists of 179 records (divided into 150 training data + 29 test data). The dataset is self-collected from CAC Bank, a financial institution providing a group of loan services in Yemen. The dataset contains of 16 text fields and 8 numeric fields.

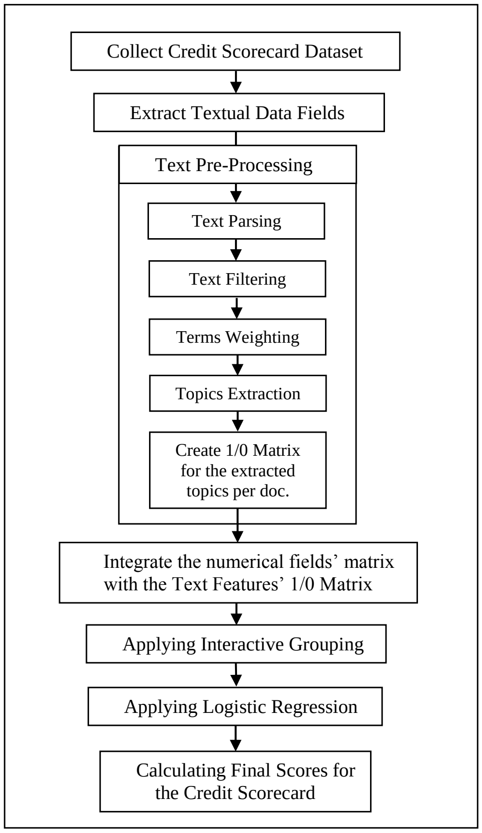
Collect Credit Scorecard Dataset
Extract Textual Data Fields
Text Pre-Processing
Text Parsing
Text Filtering
Terms Weighting
Topics Extraction
Create 1/0 Matrix for the extracted topics per doc.
Integrate the numerical fields' matrix with the Text Features' 1/0 Matrix
Applying Interactive Grouping
Applying Logistic Regression
Calculating Final Scores for the Credit Scorecard

Fig. 1: The steps that make up the proposed methodology

Four classification models were developed for the purpose of the comparison with the proposed model. Those models applied the traditional techniques of building the credit scorecard using only the numeric variables. The developed models configurations are shown in table I. The models used in the experiments are implemented using SAS Enterprise Miner tool.

TABLE I: Experimental Models Configurations

| Model | Parameter | Value |
|---|---|---|
| **Logistic Regression** | Two-Factor Interactions | N |
| | Polynomial Terms | N |
| | Regression Type | LOGISTIC |
| | Link Function | LOGIT |
| **Decision Tree** | Model Type | C4.5 |
| | Ordinal Criterion | ENTROPY |
| | Significance Level | 0.2 |
| | Maximum Branch | 2 |
| | Maximum Depth | 6 |
| **SVM** | Estimation Method | DQP |
| | Scale Predictors | Y |
| | Regularization | TUNING |
| | Constant value | 0.1 |
| | Kernel | LINEAR |
| **Neural Network** | Architecture | MLP |
| | Termination | OVERFITTING |
| | Maximum Iterations | 8 |
| | Number of Hidden Units | 2 |
| | Direct/Tanh/Sine | Y |

The performance evaluation statistics of the traditional techniques applied in building the credit scorecard by using only the numeric variables are shown in table II.

TABLE II: Classification Accuracy for Four Numerical Data Scorecard Models

| **Evaluation Method** | **Decision Tree** | **SVM** | **Neural Network** | **Logistic Regression** |
|---|---|---|---|---|
| Average Squared Error | 0.1513 | 0.193 | 0.3027 | 0.336 |
| Roc Index | 0.888 | 0.833 | 0.467 | 0.733 |
| Misclassification Rate | 0.2414 | 0.308 | 0.5172 | 0.483 |
| Wrong Classifications | 7 | 8 | 15 | 14 |

All the performance values indicate that the decision tree model outperformed the other models with the lowest average square error of ‘0.1513’ and lowest misclassification rate of ‘0.241’ which is slightly better performance than SVM. Hence, the decision tree model is used in the comparison with the proposed model. The results of the experiments are shown in table III.

TABLE III: The Evaluation Measures of The Propsoed Model

| **Evaluation Method** | **Decision Tree (Credit Scorecard Model without Text Analysis)** | **Proposed Methodology (Credit Scorecard Model with Text Analysis)** |
|---|---|---|
| True Positive | 9 | 14 |
| True Negative | 13 | 13 |
| False Positive | 1 | 1 |
| False Negative | 6 | 1 |
| Overall Accuracy (ACC) | 75.9% | 93.1% |
| Precision (PPV) | 0.90 | 0.93 |
| Recall (TPR) | 0.60 | 0.93 |
| F1 Score | 0.72 | 0.93 |

The experimental results imply that the enhancement that implemented in the credit scorecard model after applying text analysis and adding the textual variables’ extracted features to the credit scorecard is highly affecting the model’s accuracy with reference to all the statistical evaluation measures used in the comparison.

Sample of the text features in the credit scorecard resulted from applying the proposed methodology to a training data set extracted from a financial institution providing loan services in Yemen is presented in table IV.

## V. Conclusions

This paper has investigated improving the credit scorecard modelling by applying the textual data analysis for the text information filled in the forms provided by the applicants. The developed model increases the number of features in the credit scorecard by adding the textual features to the numerical features resulting from the logistic regression model after applying the pre-processing steps to the textual fields. The results of the experiments using a self-collected dataset revealed that adding the textual fields’ features improves the accuracy of the credit scorecard model by increasing the correct classification rate.

TABLE IV: Sample of the Proposed Credit Scorecard Model’s Output

| **Attribute** | **Feature** | **Score** |
|---|---|---|
| **Activity Description** | استيراد ، ماء ، مكائن تجارة ، دريلات ، كمبريوسرات | 0.226 |
| | أثاث ، صناعة الأثاث ، متخصص ، تعليمي | -0.395 |
| | missing data | -1.004 |
| | قماش ، بالجملة للأقمشة ، معتمد ، رجالي ، نسائي | 0.483 |
| | تجزئة ، تجار ، توزيع ، استهلاكية ، مورد | -0.660 |
| | جسور ، مجال الطرقات ، والجسور وصياتها ، تنفيذ مقاولات ، إنارة | 2.287 |
| | استيراد ، إستيراد الأدوية ، صيدلاني ، دواء ، مستلزم | 1.120 |
| **Main Competitor** | تابعة ، سعيد ، شركة ، هائل ، فاهم | -9.464 |
| | الحاج للدواجن ، ناصر ، مهراس ، داجنة ، الزيلعي | 0.458 |
| **Main Customers** | شركة ، بتروليم ، نكس | -0.953 |
| | مركز ، محافظة ، محلة ، وزارة الاشغال | 1.235 |
| **Main Suppliers** | تركيا ، المكلا ، ريسوت ، مؤسسة ، هائل | -0.532 |
| | الإمارات ، سيراميك ، الجزيرة ، الشارقة ، سابتكس | 0.518 |
| **Oraganization Address** | جبل ، شرق ، ذمار ، ستون ، غربي | -2.012 |
| | صناع ، مجمع ، تعز، جند ، الحوبان | 2.230 |
| | عدن ، كريتر ، فرع | -0.560 |
| **Oraganization Type** | مؤسسة ، فردة ، منشأ ، منشأة | 0.198 |
| | missing data | -5.659 |
| | حكومي ، دولة ، تابعة | 5.358 |
| | شركة ، مقفل ، مساهمة ، شركة مساهمة ، مسئولية | 0.494 |

Future studies should aim to apply other advanced statistical techniques; such as genetic algorithms and fuzzy discriminant analysis, integrated with the textual data analysis to build an enhanced credit scorecard. In addition to this, the plan is to collect larger dataset to increase the accuracy of the model.

## Acknowledgment

The authors would like to thank CAC Bank's Compliance Management for availing the dataset used in the research's experiments.

## References

[1] N. Siddiqi, Credit risk scorecards: developing and implementing intelligent credit scoring, vol. 3, John Wiley & Sons, 2012.

[2] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature", Intelligent Systems in Accounting, Finance and Management, vol.18, no. 2-3, pp. 59-88, Apr 2011.

[3] F. L. Chen and F. C. Li, "Combination of feature selection approaches with SVM in credit scoring", Expert Systems with Applications, vol. 37, no. 7, pp. 4902-4909, Jul 2010.

[4] C. L. Huang, M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, vol. 33, no. 4, pp. 847-856, Nov 2007.

[5] V. S. Desai, J. N. Crook, and G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment", European Journal of Operational Research, vol. 95, no. 1, pp. 24-37, Nov 1996.

[6] H. Abdou, J. Pointon, and A. El-masry, "Neural nets versus conventional techniques in credit scoring in Egyptian banking", Expert Systems with Applications, vol. 35, no. 3, pp. 1275-1292, Oct 2008.

[7] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", Review of Development Finance, vol. 4, no. 1, pp. 20-28, Mar 2014.

[8] F. Louzada, P. H. Ferreira-Silva, and C. A. Diniz, "On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data", Expert Systems with Applications, vol. 39, no. 9, pp. 8071-8078, Jul 2012.

[9] H. A. Abdou, "Genetic programming for credit scoring: The case of Egyptian public sector banks", Expert Systems with Applications, vol. 36, no. 9, pp. 11402-11417, Nov 2009.

[10] W. E. Henley, D. J. Hand, "A k-nearest-neighbour classifier for assessing consumer credit risk", The Statistician, pp. 77-95, Jan1996.

[11] S. Akkoc, "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data", European Journal of Operational Research, vol. 222, no. 1, pp. 168-178, Oct 2012.

[12] T. S. Lee, C. C. Chiu, C. J. Lu, and I. F. Chen, "Credit scoring using the hybrid neural discriminant technique", Expert Systems with applications, vol.23, no. 3, pp. 245-254, Oct 2002.

[13] L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", International journal of forecasting, vol. 16, no. 2, pp. 149-172, Jun 2000.

[14] I. Brown, C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications, vol. 39, no.3, pp. 3446-3453, Feb 2012.

[15] C. L. Chuang and R. H. LIN, "Constructing a reassigning credit scoring model", Expert Systems with Applications, vol. 36, no.2, pp. 1685-1694, Mar 2009.

[16] F. Thabtah, O. Gharaibeh, and H. Abdeljaber, "Comparison of rule based classification techniques for the Arabic textual data", Innovation in Information & Communication Technology (ISIICT), Fourth International Symposium on. IEEE, p. 105-111, Dec 2011.

[17] R. Feldman and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge University Press, 2007.

[18] M. Kantardzic, Data mining: concepts, models, methods, and algorithms, John Wiley & Sons, 2011.

[19] A. Mohd A Mesleh, "Chi square feature extraction based SVMs Arabic language text categorization system", Journal of Computer Science, vol. 3, no. 6, pp. 430-435, 2007.

[20] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB", Int. Arab J. e-Technol., vol. 2, no. 2, pp. 124-128, Jun 2011.

[21] R. M. Duwairi, "Arabic Text Categorization", Int. Arab J. Inf. Technol., vol. 4, no. 2, pp. 125-132, 2007.

[22] R. M. Duwairi, "A Distance-based Classifier for Arabic Text Categorization", DMIN, p. 187-192, Jun 2005.

[23] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study", DMIN, pp. 78-82, Jun 2006.

[24] R. Al-shalabi and R. Obeidat, "Improving KNN Arabic text classification with n-grams based document indexing", Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, pp. 108-112, Mar 2008.

[25] M. Al-diabat, "Arabic text categorization using classification rule mining", Applied Mathematical Sciences, vol. 6, no. 81, pp. 4033-4046, Mar 2012.

[26] M. K. Saad and W. Ashour, "Arabic text classification using decision trees", Proceedings of the 12th international workshop on computer science and information technologies CSIT, pp. 75-79, 2010.

[27] F. Harrag, E. El-qawasmeh, and P. Pichappan, "Improving Arabic text categorization using decision trees" Networked Digital Technologies, 2009. NDT'09. First International Conference on. IEEE, pp. 110-115, Jul 2009.

[28] S. Al-harbi, A. Almuhareb, A. Al-Thubaity, M.. S. Khorsheed, and A. Al-Rajeh, "Automatic Arabic text classification", 2008.

[29] A. El-Halees, "Arabic text classification using maximum entropy", The Islamic University Journal (Series of Natural Studies and Engineering), vol. 15, no. 1, pp. 157-167, 2007.

[30] A. H. Wahbeh and M. Al-kabi, "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text", Abhath Al-Yarmouk: Basic Sci. & Eng, vol.21, no. 1, pp. 15-28, 2012.

[31] F. Harrag and E. El-qawasmah, "Neural Network for Arabic text classification", Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the. IEEE, pp. 778-783, 2009.

[32] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.

[33] R. Albright, "Taming Text with the SVD", SAS Institute Inc., Cary, NC, Jan 2004.

[34] P. B. Cerrito, Introduction to data mining using SAS Enterprise Miner, SAS Publishing, 2006.

[35] D. W. Hosmer Jr and S. Lemeshow, Applied logistic regression. John Wiley & Sons, 2004.