Classification of Users' Opinions and Posts on Facebook Using Machine Learning Approaches

Ibrahim Rouby
Electronics Research Institute
Cairo, Egypt
ibrahimelgazwy@gmail.com

Mohamed Nour
Electronics Research Institute
Cairo, Egypt
mnour99@hotmail.com

Mohammed Badawy
Faculty of Electronic Engineering
Menoufia University, Menoufia, Egypt
mohamed.badawi@el-eng.menofia.edu.eg

Ehsan Abed
Electronics Research Institute
Cairo, Egypt
eabed03@yahoo.com

***Abstract*— In this research work, four classifiers are adopted, analyzed, and discussed. The classifiers are Naïve Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and Logistic Regression (LR). The classifiers are operated on a dataset with more than eight-thousands of instances. The dataset contains the users' reviews and their opinions about the quality of service of restaurants. The reviews are collected from the Arabic Facebook posts. Several experiments are done to evaluate the performance of the adopted classifiers. Moreover, some features selection methods are also applied to improve the classification process. The feature selected methods are based on term-weights with N-grams, correlation, chi-square, and mutual information. Some criteria are considered to evaluate the performance of the classification process mainly: precision, recall, F-measure, and learning time. From the experimental results, the SVM classifier outperforms the other adopted ones. Also, the feature selection method based on the correlation between the individual features and the target class outperforms the other chosen methods. The same concluding remarks are expected to take place for other datasets containing comments or reviews from social media.**

***Keywords*—** *Supervised Machine Learning, Classification Approaches, Feature Selection Methods, Facebook Reviews, and Performance Evaluation*

I. INTRODUCTION AND RELATED WORK

Text classification is one of the important themes in information retrieval and sentiment analysis. Text classification aims at identifying the class or category from a set of classes where the class text belongs to. Text classification is considered one form of supervised machine learning. Due to the large amount of text written on social media like Facebook and Twitter, the classification approaches and/or algorithms become very important in the categorization of Arabic text [1-2]. Text classification involves several steps such as document/text collection, document conversion, indexing, feature selection, training, testing, and others [3-11]. Several research efforts were presented in the literature for text classification/ categorization. This involves; but not limited to; document/ text acquisition, preprocessing operations, classification algorithms, feature selection methods, performance evaluation, and others. The terms 'text' and 'document' in this work are used interchangeably. Examples of the research efforts are briefly mentioned as shown below.

[12] presented and proposed feature ranking based on the support vector machine (SVM). The weights given by the SVM algorithm indicate the significance of those important features. The authors tested the adopted ranking features selection approach on three public datasets for text classification. The authors compared the performance of the proposed method with respect to those adopted feature selection methods. The performance of the proposed approach presented better F-measure and accuracy values. The authors in [13] presented a comparative study among the vector space model, the Naïve Bayes, and neural networks. The authors used a set of documents dedicated to all the different classes of documents. From the experimental results, the performance of the SVM outperforms the other adopted ones. The authors in [14] proposed a method for Arabic text classification. The authors compared their proposed method with three features selection metrics namely: mutual information, information gain and chi-square. The authors operated and applied the method using the SVM classifier on a dataset with more than five-thousand Arabic documents. The experimental results concluded that combining the improved method and SVM classifier outperform the performance of the other adopted methods.

[15] mentioned that logistic regression (LR) can predict the output or target in several categories. LR can explain the relation between the response variable and predictor variables. In case of sentiment analysis; for example; prediction may by positive, or negative, or neutral. The authors' work focused on some elements such as data collection, dataset labeling, data preprocessing, modeling, and performance evaluation. The authors concluded that the performance of LR using count vectorizer feature extraction was better than the corresponding results using TF-IDF. The authors in [16] presented a comparative analysis of the K-nearest neighbor (KNN), random forest (RF), and LR. The adopted classifiers were tested using the BBC news dataset. From the comparative study, the classifiers presented different and promising results for

precision, recall, accuracy, F1-measure, and confusion matrix.

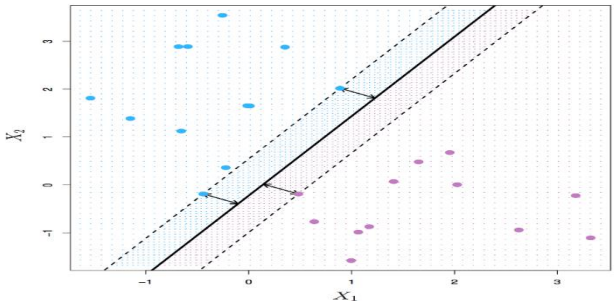
The organization of the research work will be as follows: Section 2 presents an analysis of some approaches for classifying Arabic text. The classifiers are Support Vector Machine (SVM), Naïve Bayes (NB), Stochastic Gradient Descent (SGD), and logistic regression (LR). Some feature selection methods will be discussed in section 3. Section 4; on the other hand; presents the implementation work and experimental results. Moreover, a comparative study among the performance of the adopted Arabic text classifiers is done. Finally, the discussion of results and conclusion are presented in Section 5.

II. ANALYSIS OF SOME APPROACHES FOR CLASSIFYING ARABIC TEXT

There are several types of classifiers to classify and/or categorize the different documents. The opinion mining; specifically the sentiment analysis of comments on Facebook and Twitter can be categorized using different classifiers i.e the problem of sentiment polarity classification can be handled by different classifiers. In this research work four types of classifiers will be analyzed and operated for classifying the comments on the social media. The classifiers are: support vector machine, Naïve Bayes, stochastic gradient descent, and logistic regression. Such classifiers will be presented, discussed, and applied as shown in the following subsections.

A. Classification Using Support Vector Machine

The support vector machine (SVM) is one of the common and useful classifiers. SVM is a generalization of a simple and intuitive classifier called the maximal margin classifier. SVM requires the classes to be separable by a linear boundary. The goal of the SVM is to find the optimal separating hyperplane which maximizes the margin of the training data as shown in Fig 1. There are two classes of observations as shown in Fig 1. The maximal margin hyperplane is shown as a solid line. The margin is the distance from the solid line to either of the dashed lines. The points that lie on the dashed lines are the support vectors, and the distance from those points to the margin is indicated by arrows. The points indicate the decision rule made by a classifier based on this separating hyperplane [17]. It is assumed that a data matrix X is given. The data matrix dimensions are nxp where n is the number of training observations in the p-dimensional



space.

Fig. 1: The Maximal Margin Hyperplane [12], and [18]

$$x_1 = \begin{pmatrix} x_{11} \\ \cdot \\ \cdot \\ \cdot \\ x_{1p} \end{pmatrix}, \ldots, x_n = \begin{pmatrix} x_{n1} \\ \cdot \\ \cdot \\ \cdot \\ x_{np} \end{pmatrix} \quad (1)$$

The observations fall into two classes that is: $y_1, \ldots\ldots y_n \in \{-1,1\}$ where "-1" and "1" represent respectively the negative class and positive classes. A p-vector of observed features $x^* = (x_1, x_2, \ldots x_p)^T$. The goal is to develop a classifier based on the training data that will correctly classify the test observation using its feature measurements. Then, a separating hyperplane has the property that

$$\beta_0 + \beta_{1xi1} + \beta_{2xi2} + \cdots + \beta_{pxip} > 0 \; if \; y_i = 1, \quad (2)$$

And

$$\beta_0 + \beta_{1xi1} + \beta_{2xi2} + \cdots + \beta_{pxip} > 0 \; if \; y_i = -1, \quad (3)$$

Equivalently, a separating hyperplane has the property that

$$y_i(\beta_0 + \beta_{1xi1} + \beta_{2xi2} + \cdots + \beta_{pxip}) \quad (4)$$

If $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients of the maximal margin hyperplane, then the maximal margin classifier classifies the test observation x* based on the sign of $f(x*) = \beta 0 + \beta 1x*1 + \beta 2x*2 + \ldots + \beta px*p$ for all i=1, 2,…n. If(x*) is positive, then the test observation is assigned to class "1", and if f(x*) is negative, it is assigned to class "-1".
Moreover, the task of constructing the maximal margin hyperplane is considered based on a set of n training observations $x_1, x_2, \ldots x_n \in R^p$ where the associated class labels $y_1, y_2, \ldots y_n \in \{-1,1\}$. Briefly, the maximal margin hyperplane is the solution to the optimization problem. For more details about SVM, the reader can refer to [١٧], and [18].

B. Classification Using Naïve Bayes

Naïve Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "**Naïve**". The naïve Bayesian classifier, or simple Bayesian classifier, works as follows [19]:

1. Let **D** be a training set of tuples and their associated class labels. As usual, each tuple is represented by an **n-dimensional** attribute vector, **X = (x_1, x_2, x3, …, x_n)**, depicting **n** measurements made on the tuple from n attributes, respectively, **A_1, A_2, … , A_n.**
2. Suppose that there are **m** classes, **C_1, C_2, …, C_m.** Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \; for \; 1 < j \leq m, j \neq 1 \quad (5)$$

Thus, it is required to maximize $P(C_i|X)$. The class C_i for which $P(C_i|X)$ is maximized and it is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)} \qquad (6)$$

3. As P(X) is constant for all classes, only $P(C_i|X)P(C_i)$ needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2)=\ldots=P(C_m)$, and therefore $P(X|C_i)$ is maximized. $P(X|C_i)P(Ci)$ is maximized such that the class prior probabilities may be estimated by $P(C_i) = |C_{i,D}|/|D|$ where $|C_{i,D}|$ is the number of training tuples of class C_i in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. To reduce computation in evaluating $P(X|C_i)$, the Naive assumption of class-conditional independence is made. This presumes that the attributes' values are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$
$$= P(x_1|C_i) * P(x_2|C_i) * \ldots * P(x_n|C_i) \qquad (7)$$

It is easy to estimate the probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, …, $P(x_n|C_i)$ from the training tuples. For more details the reader can refer to [19], and [٢0].

C. Classification Using Stochastic Gradient Descent

Optimization is important when dealing with a problem like building software products. Stochastic Gradient Descent (SGD) is a simple approach to discriminative learning of linear classifiers under convex loss functions. Gradient descent is a way to minimize an objective function $J(\theta)$ parameterized by a model's parameter $\theta \in R^d$ by updating the parameters in the opposite direction of the gradient of the objective function $\nabla_\theta J(\theta)$ w.r.t. the parameters. The learning rate **η** determines the size of the steps taken to reach a local minimum. In other words, the slope of the surface created by the objective function downhill is followed until a valley is reached. The SGD approach performs a parameter update for each training example $\mathbf{x^{(i)}}$ and label $\mathbf{y^{(i)}}$.

$$\theta = \theta - \eta.\nabla_\theta J(\theta; x^{(i)}; y^{(i)}) \qquad (8)$$

Batch gradient descent performs redundant computations for large datasets, as it recomputes gradients for similar examples before each parameter update. SGD can make one update at a time. SGD can also perform frequent updates with a high variance that causes the objective function to fluctuate heavily as shown in Figure 2.

The stochastic gradient descent (SGD) can be presented as follows:

- Choose an initial vector of parameters **θ** and learning rate **η**.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle examples in the training set.
  - For i=1, 2, ..., n do:
    $\theta = \theta - \eta.\nabla_\theta J(\theta; x^{(i)}; y^{(i)})$

For more details about SGD the reader can refer to [21].

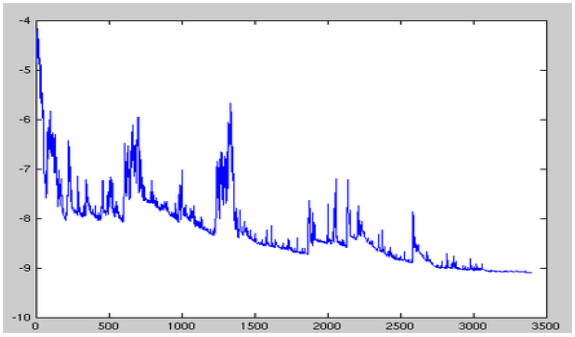


Fig. 2: Fluctuations in the Objective Function as Gradient Steps w.r.t. Mini-Batches [21], and [22]

D. Classification Using Logistic Regression

Logistic regression (LR) is one of the machine learning algorithms used for solving the classification problem and predicting the classes. LR can predict the probability of occurrence of an event utilizing a logistic function.
[23] mentioned that LR is a special case of linear regression where the target variable is categorical. In LR, the dependent variable can follow the Bernoulli distribution and estimation can be done through the maximum likelihood. LR may be binary, multinomial or ordinal. In binary LR; the target variable has only two possible outcomes where in multinomial LR; it has three or more nominal categories such as predicting the type of an object. The target variable in ordinal LR has three or more ordinal categories such as product rating from one to five.

LR is a supervised learning algorithm used in classifying the individuals in the categories based on logistic function. The mathematical version of LR can be briefly presented by beginning with a simple linear regression and applying the sigmoid function. The formula of the simple linear regression can be represented by:

$$y = b_0 + b_1 * x \qquad (9)$$

The sigmoid function formula can be briefly represented by

$$P = \frac{1}{1 + e^{y}} \qquad (10)$$

Where by substitution, the LR formula can be written as

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 * x \qquad (11)$$

$$\text{Lagit}(S) = b_0 + b_1 m_1 + b_2 m_2 + b_3 m_3 + \cdots b_k m_k \qquad (12)$$

Where S is the probability of the presence of interest features, m_1, m_2, m_3, ……. m_k ,m-1.„m-2.,...,m-k.are the predictor values and b_0, b_1, b_2, b_3, ….b_k are the intercept of the classifier. For more details about the LR algorithm, the reader can refer to [15-16] and [24].

III. ANALYSIS OF SOME FEATURE SELECTION METHODS

Feature selection approaches are important in machine learning. Using n- gram language models for text classification may lead to high dimensional datasets. Machine learning algorithms need a different form of input to make it possible to compute satisfying results. The input for a machine learning algorithm is represented as a vector of weighted features. The feature is defined as a string within a document. Furthermore, the process of turning a corpus into numerical feature vectors is called vectorization. Each feature in a document is assigned a weight. There are several approaches for features extraction. The approaches such as: Term Frequency-Inverse Document Frequency (TF-IDF) with n-grams, correlation, mutual information, and chi-squared are adopted in this work.

A. Feature Selection Method Based on Term-Weighting and N-Grams

TF-IDF is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in the collection. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the collection. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. The TF-IDF weight is composed of two terms: TF and IDF. The normalized Term Frequency (TF) is the number of times a word appears in a document divided by the total number of words in that document. The Inverse Document Frequency (IDF) is the logarithm of the number of documents in the collection divided by the number of documents where the specific term appears [25], and [26].

$$W_{x,y} = TF_{x,y} * \log(N/DF_x) \qquad (13)$$

where

$W_{x,y}$: The weight of term x in document y,
$TF_{x,y}$: The frequency of term x in document y,
N: The total number of documents, and
DF_x: The number of documents containing the term x.

Regarding TF-IDF, it is important to consider the following:

- TF-IDF is a simple model that is expected to present great results.
- TF-IDF features creation is a fast process, which will lead to shorter waiting time.
- The feature creation process is better to avoid issues like overfitting.

For more details the reader can refer to [25-27].

However, important details about the original document such as phrases, word order, context and sentences are lost. Alternatively, adding multi-word expressions can be helpful identifying certain multi-word expressions, such as "United Kingdom" or "white house". N-grams are basically sequences of n consecutive words from a given text. For example, considering the following sentence: "My favorite treat is cheeseLake", would create the following n-grams:

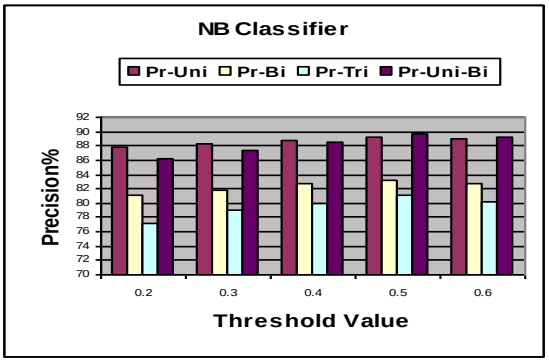
TABLE I: N-GRAMS FEATURE REPRESENTATION

Uni-grams (n=1)	My	Favorite	treat	Is	cheeseLake	
Bi-grams(n=2)	My favorite		favorite treat	Treat is	is cheeseLake	
Tri-grames(n=3)	My favorite treat			favorite treat is		treat is cheeseLake

When creating the features with this method, parameters can be chosen.

- N-grams range: uni-grams, bi-grams, and tri-grams are considered.
- Maximum/Minimum Document Frequency: when building the vocabulary, the terms that have a document frequency strictly higher or lower than the given threshold are ignored.
- Maximum features: the required features ordered by term frequency across the corpus are chosen.

Moreover, the classifiers NB, SVM, SGD, and LR as well as the feature selection methods are operated and applied on the dataset mentioned in Section 4. By changing the threshold value (weight value), the number of selected features will be also changed. Also, the experimental results will be changed by charging the number of selected features. The values of measurable criteria used for evaluating the classification process will change depending on the adopted classifiers as well as the feature selection approaches. The reported measurable criteria; in this work; are precision, recall, F-measure, and learning time. Figures 3.a, 3.b, 3.c and 3.d show respectively the experimental results of precision, recall, F-measure, and learning time for applying the uni-gram, bi-gram, tri-gram, and combination of both the uni-gram and bi-gram. Such figures are reported for the NB classifier. Similarly, the same experiments are also applied and operated on the



SGD, LR, and SVM classifiers as shown respectively in Figures 4.a to 4.d, 5.a to 5.d, and 6.a to 6.d.

Fig. 3.a: Precision for NB Classifier

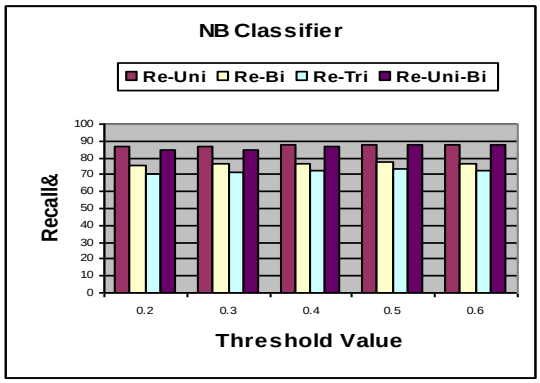


Fig. 3.b: Recall for NB Classifier

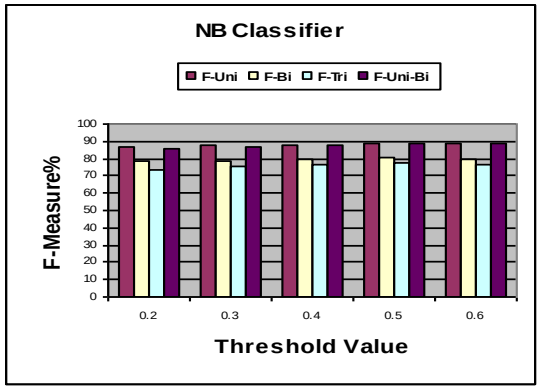


Fig. 3.c: F-measure for NB Classifier

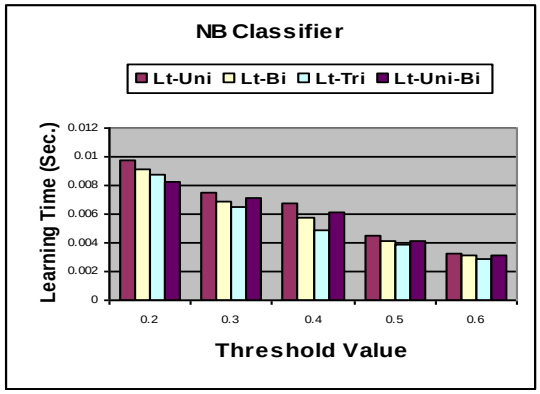


Fig. 3.d: Learning Time for NB Classifier

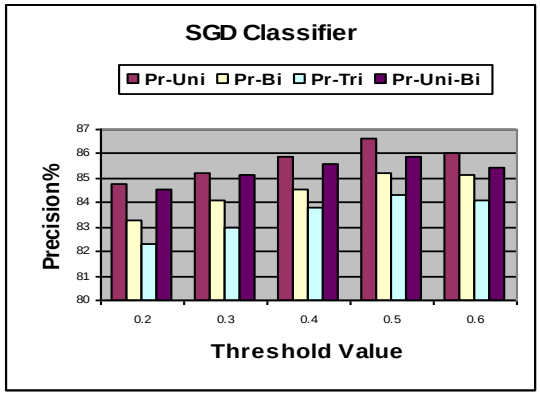


Fig. 4.a: Precision for SGD Classifier

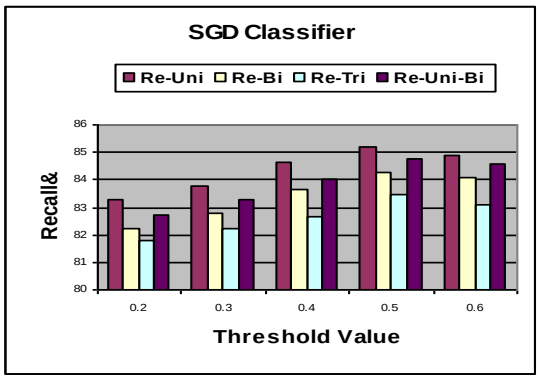


Fig. 4.b: Recall for SGD Classifier

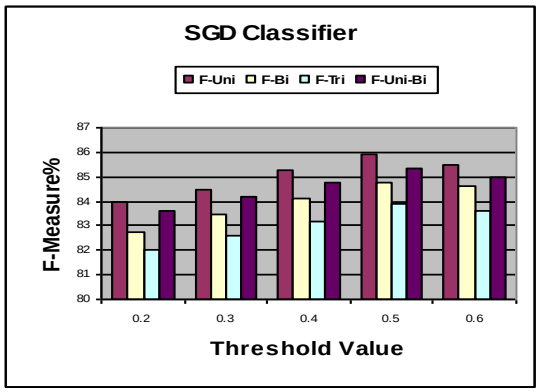


Fig. 4.c: F-measure for SGD Classifier

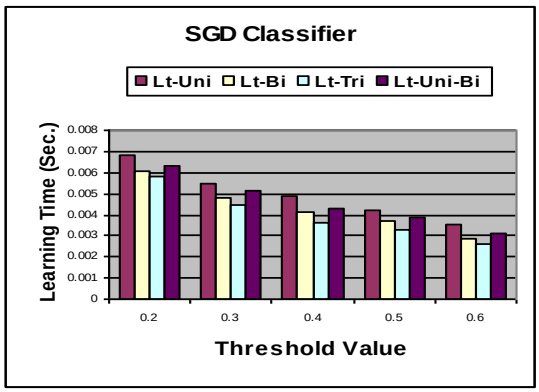


Fig. 4.d: Learning Time for SGD Classifier

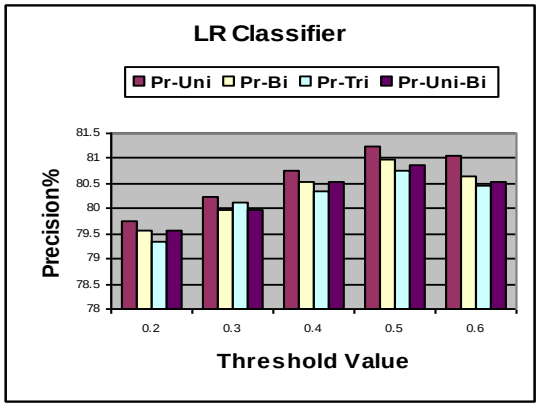


Fig. 5.a Precision for LR Classifier

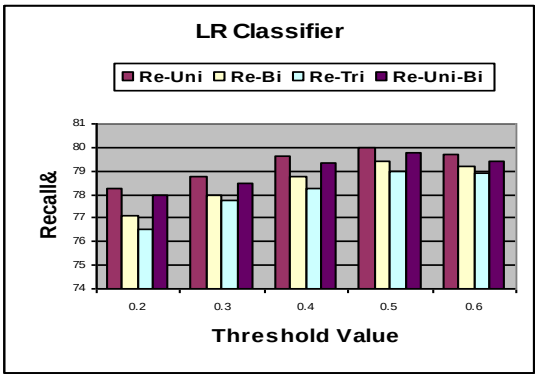


Fig. 5.b: Recall for LR Classifier

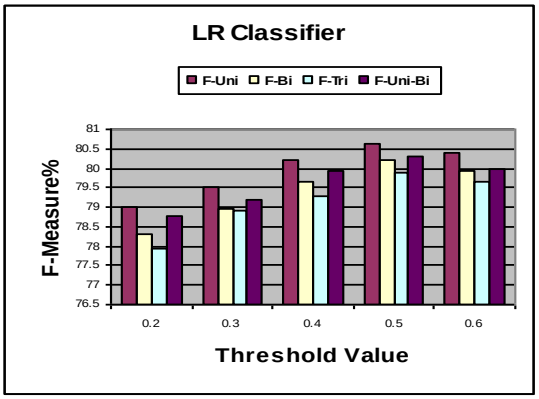


Fig. 5.c: F-measure for LR Classifier

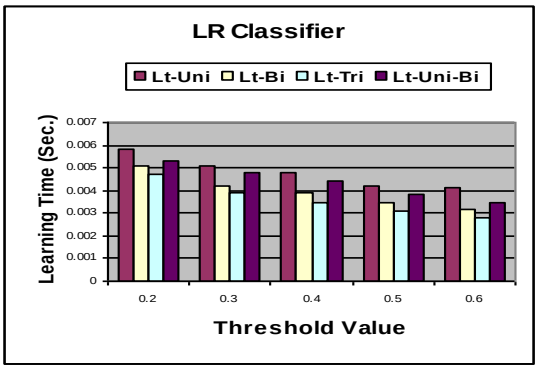


Fig. 5.d: Learning Time for LR Classifier

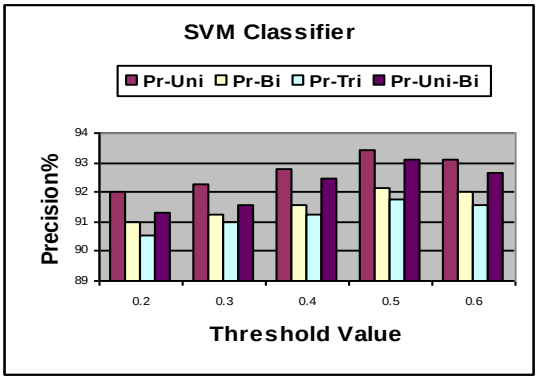


Fig. 6.a: Precision for SVM Classifier

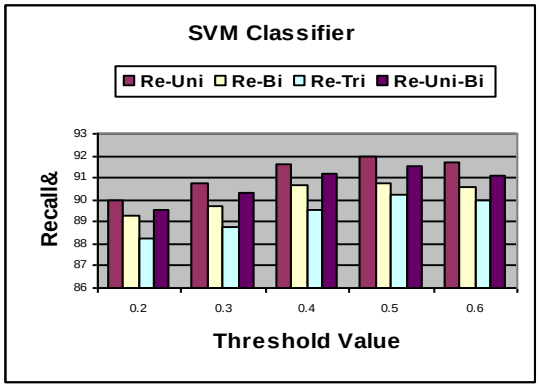


Fig. 6.b: Recall for SVM Classifier

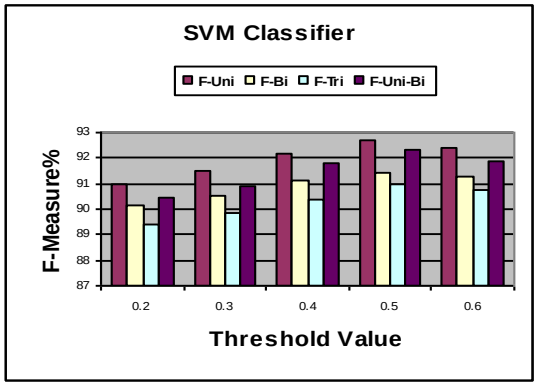


Fig. 6.c: F-measure for SVM Classifier

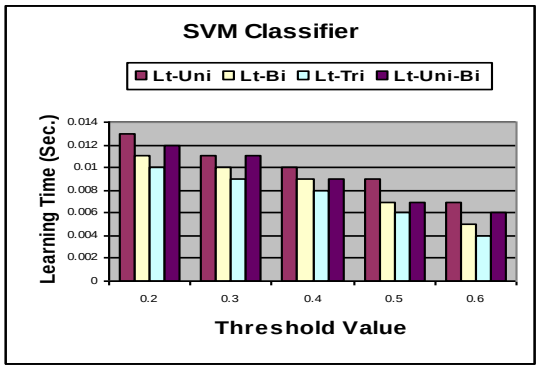


Fig. 6.d: Learning Time for SVM Classifier

B. Feature Selection Method Based on Correlation between Individual Features and Target Classes

Feature selection is important to reduce the dimensionality of the text feature space. Using a feature selection method, the dimension of the space is reduced by selecting the most significant features. The correlation between individual features and target classes is considered as a statistical analysis approach over the feature space. This is necessary to select a discriminative subset of features and/or the most significant ones. This approach aims to find the relationship between every individual feature and the target variable. Correlation R(i) between any feature vector x_i and the class vector **y** can be computed using the formula shown in equation 14. Each feature will have a test score and/or a correlation value.

The features with top score or high correlation values are selected.

$$R(i) = \frac{\sum_{k=1}^{m}(x_{k,i} - x_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{k,i} - x_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}} \quad (14)$$

Where x_i, and **y** are respectively a feature vector and a target or class vector while m is the number of instances [٢٨]. The experimental results of precision, recall, F-measure, and learning time for applying the correlation method are shown respectively in Figures 3.e, 3.f, 3.g, and 3.h. Such figures are reported for the NB classifier. Similarly, the same experiments are also operated for the SGD, LR, and SVM classifiers as shown respectively in Figures 4.e to 4.h, 5.e to 5.h, and 6.e to 6.h.

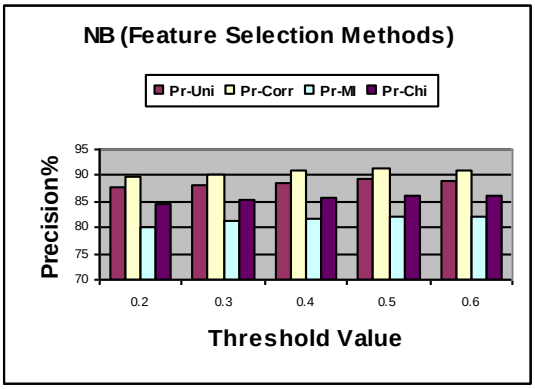


Fig. 3.e: Precision for Feature Sel. Methods

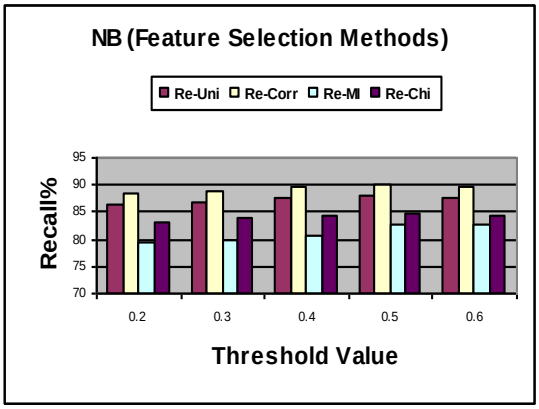


Fig. 3.f: Recall for Feature Sel. Methods

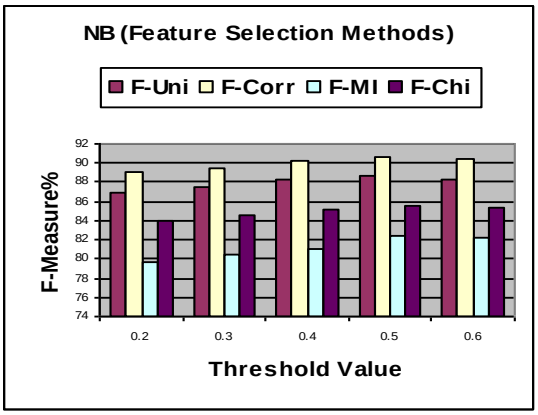


Fig. 3.g: F-measure for Feature Sel. Methods

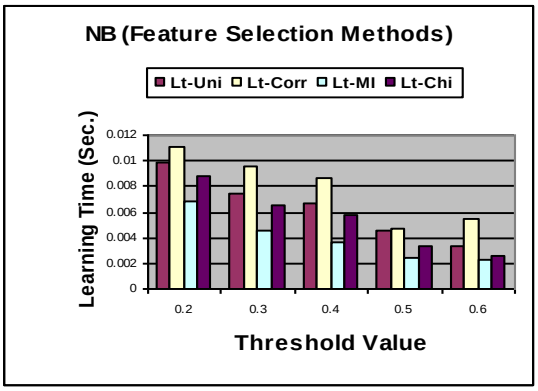


Fig. 3.h: Learning Time for Feature Sel. Methods

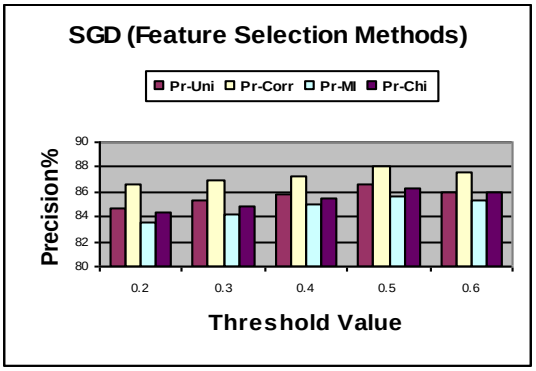


Fig. 4.e: Precision for Feature Sel. Methods

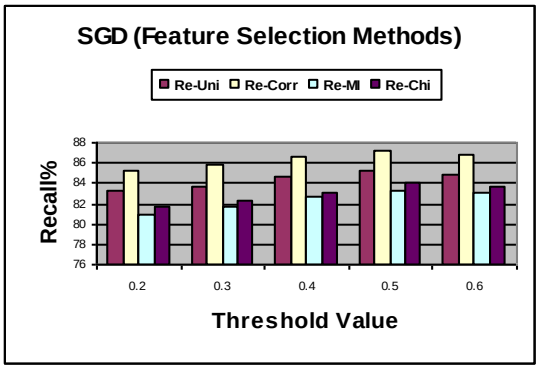


Fig. 4.f: Recall for Feature Sel. Methods

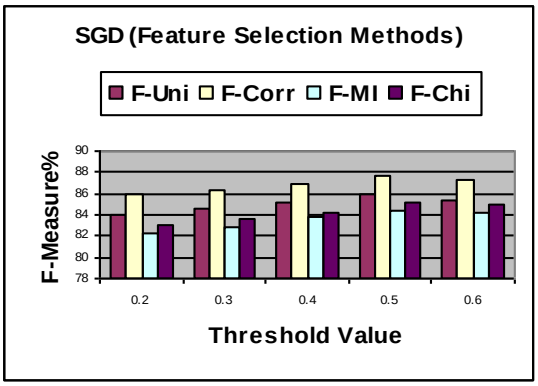


Fig. 4.g: F-measure for Feature Sel. Methods

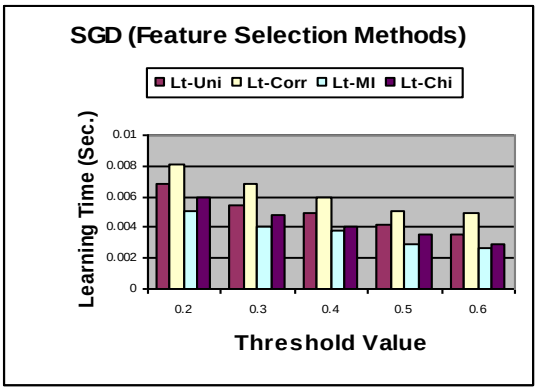


Fig. 4.h: Learning Time for Feature Sel. Methods

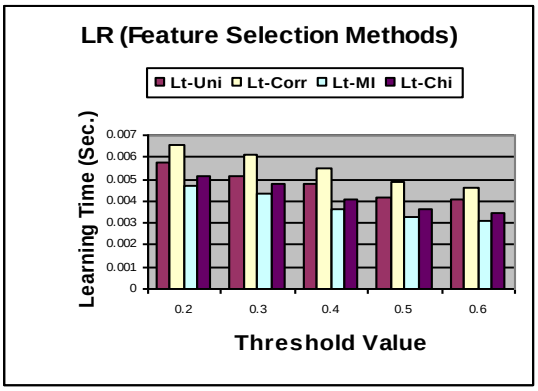


Fig. 5.h: Learning Time for Feature Sel. Methods

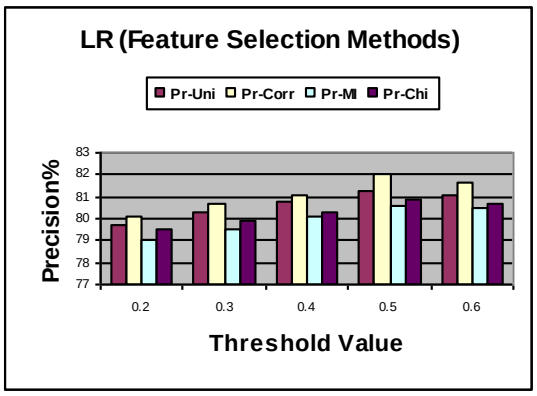


Fig. 5.e: Precision for Feature Sel. Methods

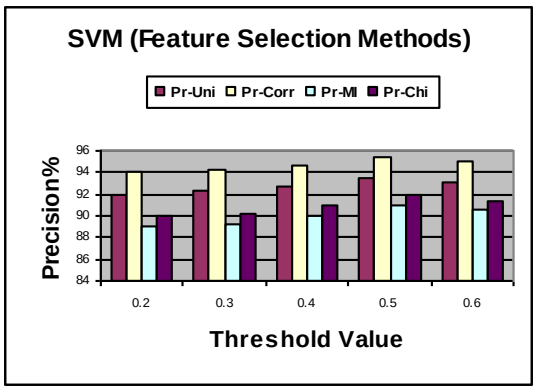


Fig. 6.e: Precision for Feature Sel. Methods

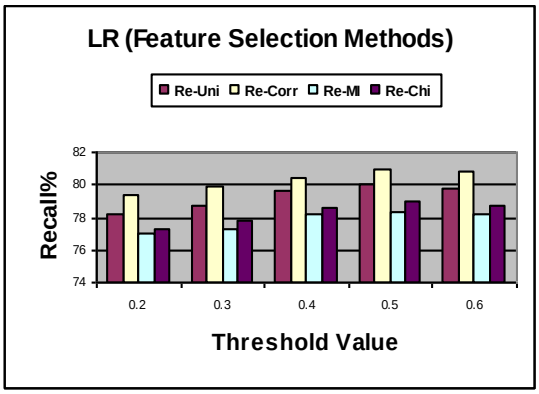


Fig. 5.f: Recall for Feature Sel. Methods

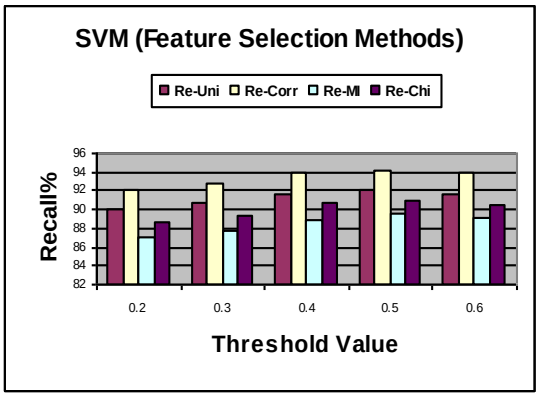


Fig. 6.f: Recall for Feature Sel. Methods

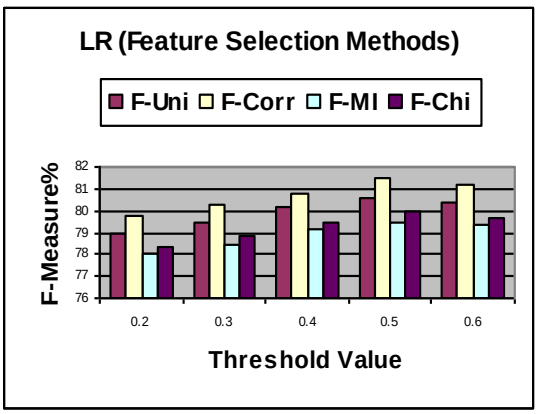


Fig. 5.g: F-measure for Feature Sel. Methods

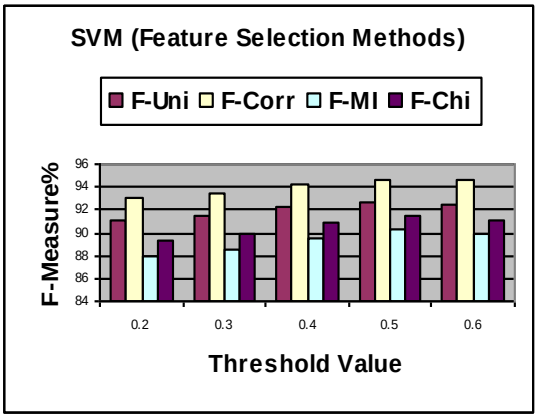


Fig. 6.g: F-measure for Feature Sel. Methods

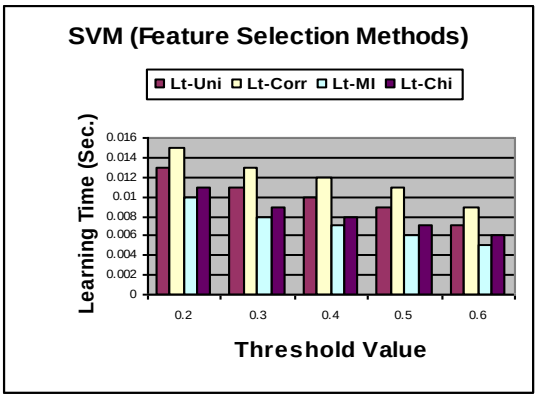


Fig. 6.h: Learning Time for Feature Sel. Methods

C. Feature Selection Method Based on Mutual Information

Mutual information (MI) is used to measure the dependency between the variables. MI equals zero if and only if two random variables are independent, and higher values mean higher dependency. MI between two random variables is a non-negative value. MI is a measure of association between variables, capturing both linear and non-linear dependencies that have gained wide acceptance [29-30]. MI between two discrete random variables X and Y, denoted MI(*X*, *Y*) is defined by:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} P_{XY}(x,y) \ln \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} \qquad (15)$$

where $P_{XY}(x,y)$ is the joint probability distribution, and $P_X(x)$ and $P_Y(y)$ are the marginal represented by the following equations:

$$P_X(x) = \sum_y P_{XY}(x,y) \qquad (16)$$

$$P_Y(y) = \sum_x P_{XY}(x,y) \qquad (17)$$

The experimental results of precision, recall, F-measure, and learning time for applying the mutual information method are shown respectively in Figures 3.e, 3.f, 3.g, and 3.h. Such figures are reported for the NB classifier. Similarly, the same experiments are also operated for the SGD, LR, and SVM classifiers as shown respectively in Figures 4.e to 4.h, 5.e to 5.h, and 6.e to 6.h.

D. Feature Selection Method Based on Chi-squared

Chi-squared (CHI) is a supervised, one-sided feature selection method that calculates the correlation of term **t** with class **C** [30-33] . CHI is calculated as:

$$\mathrm{CHI}_{(t,c)} = \frac{D * (PN - MQ)^2}{(P+M) * (Q+N) * (P+Q) * (M+N)} \qquad (18)$$

Where
D: Total number of documents,
P: Number of documents of class ‘c’ containing the term ‘t’, *Q :* Number of documents containing ‘t’ occurring without ‘c’, *M:* Number of documents class ‘c’ occurring without ‘t, and *N*: Number of documents of other classes without ‘t’.

The experimental results of precision, recall, F-measure, and learning time for applying the chi-squared method are shown respectively in Figures 3.e, 3.f, 3.g, and 3.h. Such figures are reported for the NB classifier. Similarly, the same experiments are also operated for the SGD, LR, and SVM classifiers as shown respectively in Figures 4.e to 4.h, 5.e to 5.h, and 6.e to 6.h.

IV. IMPLEMENTATION WORK

In this section, a set of experiments are presented to apply the adopted classification algorithms and the chosen feature selection methods. To evaluate the effectiveness of the algorithms and feature selection methods, a dataset or a document collection is taken as a testbed. The dataset is partitioned into two parts: training and testing. The selection of training and testing sets is randomized while the number of instances of training set is always greater than double of that number dedicated for testing. A comparative study is done among the behavior of the adopted classifiers as well as the chosen feature selection methods. In this concern, the percentage values of precision, recall, F-measure, and learning time are shown respectively in Figures 7.a to 7.h.

Moreover and before applying the classifiers on the dataset, a preprocessing operation is done to clean the dataset. Stemming and removing of stopwords, repeated words, and special characters are done to simplify the extraction of the features. The main characteristics of the chosen dataset and adopted measurable criteria are briefly mentioned in the following subsections.

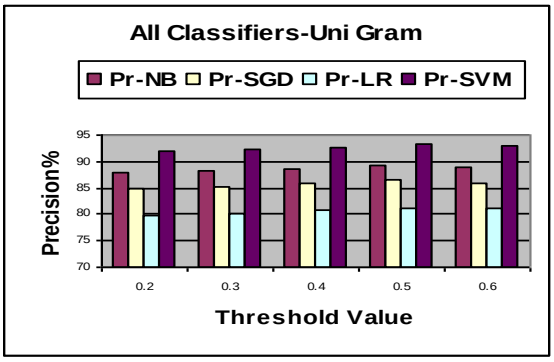


Fig. 7.a: Precision for All Classifiers

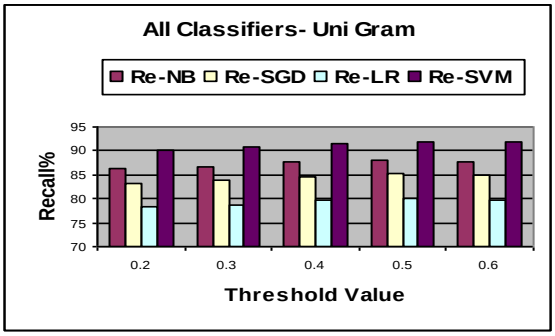


Fig. 7.b: Recall for All Classifiers

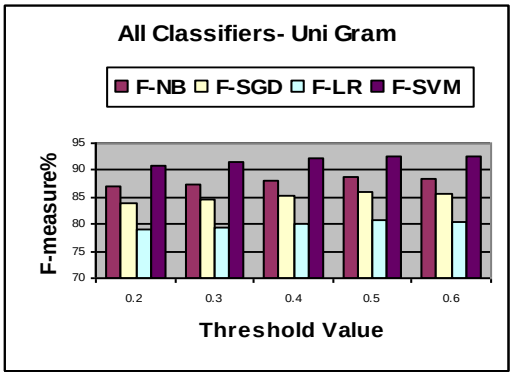


Fig. 7.c: F-measure for All Classifiers

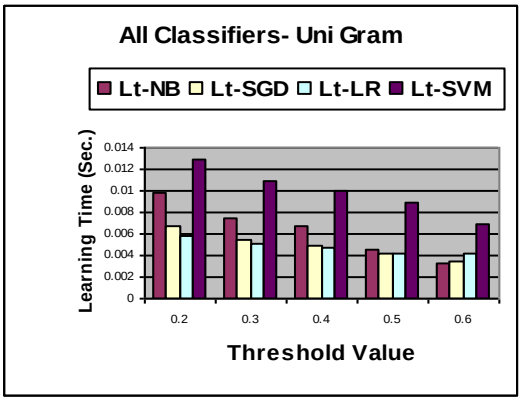


Fig. 7.d: Learning Time for All Classifiers

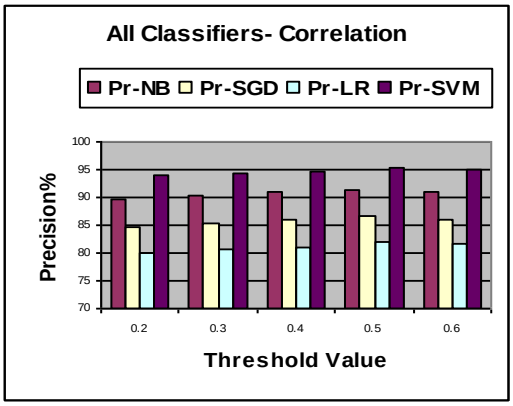


Fig. 7.e: Precision for All Classifiers

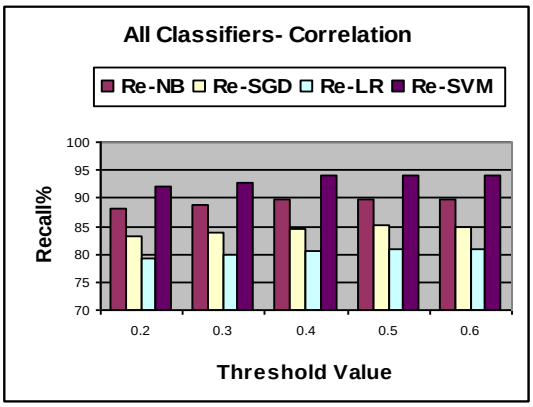


Fig. 7.f: Recall for All Classifiers

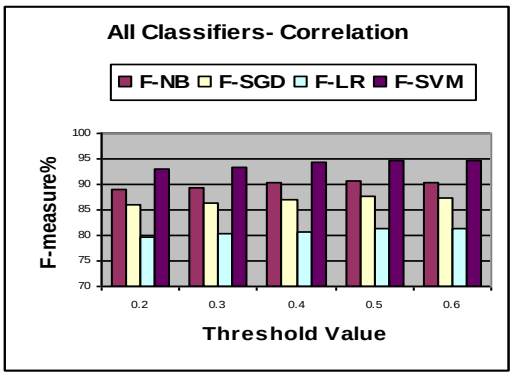


Fig. 7.g: F-measure for All Classifiers

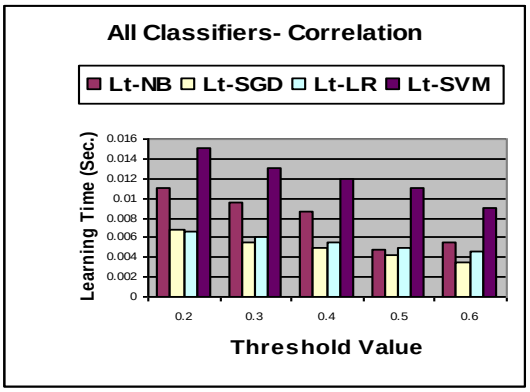


Fig. 7.h: Learning Time for All Classifiers

A. The Characteristics of the Chosen Dataset Collection

A dataset is considered the input to the preprocessing operation. The dataset presents the users' reviews about restaurant customers. The dataset was collected from the Facebook [34]. The descriptive information and/or characteristics of the dataset are illustrated in Table 2.

TABLE 1: DESCRIPTIVE INFORMATION OF THE CHOSEN DATASET [٢4]

Dataset Name	Reviews of Restaurant Customers
Total Number of Reviews	8341
Number of Positive Reviews	2413
Number of Negative Reviews	5928
Number of Vocabularies	27497
Average Number of Tokens/Review	19
Number of Classes	2

B. Evaluation Measurable Criteria

Precision, recall, F-measure, and learning time(Sec.) are considered standard measures used in text mining. The criteria are briefly mentioned as shown in equations 13 to 16 respectively [28] and [35-37].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (19)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (20)$$

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (21)$$

where TP, TN, FP, and FN represent the number of true positive, number of true negative, number of false positive, and number of false negative respectively. Classification accuracy is the proportion of true positive and true negative obtained by the algorithms over the total number of instances. Precision is the proportion of the true positive against the true positive and false positive. Recall is the proportion of the true positive against the true positive and false negative. Finally, the F-measure is a relationship between the precision and recall and can be computed using equation (21) [٢٨].

V. DISCUSSION OF RESULTS AND CONCLUSION

In this work, four classification approaches and four feature selection methods were adopted, analyzed, implemented, and evaluated. The classifiers and the feature selection methods were applied on a dataset containing reviews about restaurants collected from the Facebook. The dataset contains more than eight-thousands of instances or reviews.

From the experimental results it is easy to say that precision%, recall %, F-measure%, and learning time (Sec.) were different for the four classifiers. Also, the values of the measurable criteria were different for the same classifier for the different adopted feature selection methods. In the experiments, threshold values of the feature weights were taken into consideration. Figures 3.a to 3.d show respectively the precision%, recall %, F-measure%, and learning time (in seconds) for the behavior of the NB classifier using the uni-gram, bi-gram, tri-gram, and combined uni-gram with bi-gram. The values changed by changing the threshold value. The learning time (in Sec.) was also different for the different threshold values. This is due to the different values of the number of selected features. Figure 3.e to 3.h presented the precision%, recall%, F-measure%, and learning time (in Sec.) for applying the NB classifier on the adopted four feature selection methods. The experimental results were different for the different threshold values and also for the different feature selection methods.

Similarly, Figures 4.a to 4.d, 5.a to 5.d, and 6.a to 6.d show respectively the precision%, recall%, F-measure%, and learning time for the SGD, LR, and SVM classifiers. Figures 4.e to 4.h, 5.e to 5.h, and 6.e to 6.h show also the performance of the same measurable criteria for applying the SGD, LR, and SVM classifiers on the four feature selection methods respectively.

From the experimental results shown in the Figures mentioned above it was noticed that the precision%, recall%, and F-measure% were better for the uni-gram than those corresponding values of the bi-gram, and tri-gram. The worst performance values were for using the tri-gram. Moreover, sometimes the combination of both uni-gram and bi-gram present good experimental results very close to those values obtained by the uni-gram. This occurred for the four classifiers as shown respectively in Figures 3.a to 3.c, 4.a to 4.c, 5.a to 5.c, and 6.a to 6c. The learning time (in Sec.); on the other hand; was the smallest when using the tri-gram while the highest values of learning time were for the unigram. This occurred for the whole experiments as shown in Figures (3.d, 3.h), (4.d, 4.h), (5.d, 5.h), and (6.d, 6.h) for the NB, SGD, LR, and SVM classifiers respectively. In all cases, the learning time (in Sec.) was different for the different threshold values for all classifiers using the same test-bed dataset.

Moreover, a set of experiments were done using the adopted feature selection methods. It was noticed that the values of the measurable criteria were different for each feature selection method. By changing the threshold value, the number of selected features was also changed. It was also noticed in the majority of experiments that the values of the measurable criteria were not good for using the more number of features. The precision%, recall%, and F-measure% become better by decreasing the number of features till a certain value then the performance of the measurable values become to decrease. The best performance occurred for the number of features equals 600 in our experiments. The experimental results for applying the adopted feature selection methods are shown respectively in Figures 3.e to 3.h, 4.e to 4.h, 5.e to 5.h, and 6.e to 6.h for the NB, SGD, LR and SVM classifiers. Comparing the values of the obtained results for the feature selection methods, it was noticed that the best performance occurred for that method based on the correlation between the individual features and the target classes. The worst performance values were for the feature selection method based on mutual information.

Generally speaking and from the set of operated experiments it is easy to say that the best performance is for the SVM classifier while the worst behavior is for the SGD classifier. The performance of the uni-gram; on the other hand; is better than those corresponding values of bi-gram, and tri-gram respectively. The performance of the feature selection method based on correlation is the best while that one based on the mutual information is the worst compared with the other adopted feature selection methods. The same concluding remarks are expected to take place for using other test-beds containing reviews or comments from the social media.

REFERENCES

[1] Musab Hijazi, Akram Zaki, and Amelia Ismail, "Arabic Text Classification: Review Study", Journal of Engineering and Applied Sciences, Vol. 11, No. 3, PP. 528-536, 2016.

[2] Mavy Al-Tahrawi and Sumaya Al-Khtib, "Arabic Text Classification using Polynomial Networks", Journal of King Saud University-Computer and Information Science, Vol .27, PP. 437- 449, 2015.

[3] Mehdi Allahyari, Seyedamin Powiyeh, Mehdi Assefi, Saied Safaei, Elizabeth Trippe, and Juan Gutierez, "A Brief Survey of Text Mining: Classification, Clustering, and Extraction Techniques", Journal of KDD Bigdas, Halifax, Canada, PP. 1-13, Jully 2017.

[4] Aaditya Jain and Jyati Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification", The International Journal of Computer Application, Vol. 6, No. 2, PP. 126-129, April 2016.

[5] Ashraf Odeh, Ayrnan Abu-Errub, Qusai Shambowr, and Nidal Turahb, "Arabic Text Categorization Algorithm Using Vector Evaluation Method", The International Journal of Computer Science and Information Technology (IJCSIT), Vol. 6, No. 6, PP. 83-92, December 2014.

[6] Rutviya Pandya, and Jayati Pandya, "C5.0 Algorithm to Improve Decision Tree with Feature Selection and Reduced Error Pruning", The International Journal of Computer Applications, Vol. 117, No. 16, PP. 18-21, May 2015.

[7] Rami Ayadi, Mohsen Maraoui, and Mounir Zrigui, "A Survey of Arabic Text Representation and Classification Methods", The Journal of Research in Computing Science, Vol. 117, PP. 51-62, 2016.

[8] Ching-Hue Cheng, and Khaidou Zrcik, "Sentimental Text Mining Based on an Additional Features Method for Text Classification", PLOS ONE, Vol. 14, No. 6, PP. 1-17, 2019.

[9] A. S. Galathiya, A. P. Ganatra, and C. K. Bhencsdadia, "Classification with an Improved Decision Tree Algorithm", International Journal of Computer Applications, Vol. 46, No. 23, PP. 1-6, May 2012.

[10] Mandeep Choudhary and V. S. Dhaka, "Automatic e-mails Classification Using Genetic Algorithms", The International Journal of Computer and Information Technologies, Vol, 6, No. 6, PP. 5097-5103, 2015.

[11] Sing-Sarn Hong, Wanhee Lee, and Myung-Moak Han, "The Feature Selection Method Based on Genetic Algorithm for Efficient of Text Clustering and Text Classification", The International Journal of Advanced Soft Computing and Applications, Vol. 7, No. 1, PP. 22-40, March 2015.

[12] Thabit Sabbah, Mosab Ayyash, and Mahmoud Ashraf, "Support Vector Machine Based on Feature Selection Method for Text Classification", The International IEEE Arab Conference on Information Technology, Yassmine Hammam, Tunisia, PP. 1-8، December 22-24, 2017.

[13] Adel Hamdah Mohammad, Tasiq Alwada, and Omar Al-Momari, "Arabic Text Categorization Using Support Vector Machine, Naïve Bayes, and Neural Network", The GSTF Journal on Computers (JOC), Vol. 5, No. 1, PP. 108-115, August 2016.

[14] Said Bahassine, Abdallah Madani, Mohammed Al-Sarem, and Mohamed Kissi, "Feature Selection Using an Improved Chi-Square For Arabic Text Classification", Journal of King Saud University-Computer and Information Sciences, Vol. 32, Issue 2, PP. 225-231, 2020.

[15] Umniy Salamah, and Desi Ramayanti, "Implementation of Logistic Regression Algorithm for Complaint Text Classification in Indonesian Ministry of Marine and Fisheries", The International Journal of Computer Techniques, Vol. 5, Issue. 5, PP. 74-78, September-October 2018.

[16] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification", The Augmented Human Research, Springer, Vol. 5, No. 12, PP. 1-16, 2020.

[17] Gareth James, Daniela Witten, Tervor Hastie, and Robert Tibshirani, "An Introduction to Statistical Learning with Applications in R", Springer Text in statistics, SBN 978-1-4614-7138-7, June 24, 2013.

[18] Maganti Syamaaha, N. J. Nalini, Lafeshamaera Phaneendra Maguluri, and R. Ragupatly, "Comparative Analysis of Document Level Text Classification Algorithms Using R", The International TOP Conference Series: Material Science and Engineering 225, Narsimha Reddy Engineering College, India, PP. 1-8, 2017.

[19] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Concepts and Techniques", Elsevier, 2012.

[20] Feng-Jen Yang, "An Implementation of Naive Bayes Classifier", The International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, Nevada, USA, 2018.

[21] Tim Jahn, and Bangti Jin, "On the Discrepancy Principle for Stochastic Gradient Descent", arXiv:2004.14625V1[math.NA], 2020, https://arxiv.org/ abs/2004.14625.

[22] "Stochastic Gradient Descent" Downloaded From the Internet From the Website https://en.wikipedia.org/wiki/Stochastic_gradient _descent.

[23] "Understanding Logistic Regression in Python", Downloaded From the Internet From the Website https://www.datacamp.com/ community/tutorials/ understanding-logistic-regression-python, 2020.

[24] Fahad Mazaed Alotaibi, "Classifying Text-Based Emotions Using Logistic Regression", VFAST Transactions on Computer Sciences, Vol. 7, No. 1, PP. 31-37, January-December 2019.

[25] H. Wu, R. Luk, K. Wong and K. Kwok, "Interpreting TF-IDF Term Weights as Making Relevance Decisions", ACM Transactions on Information System, Vol. 26, No. 3, Article 13, PP. 1-37, 2008.

[26] Fu Yu, and Yu You, "Research on Text Representation Method Based on Improved TF-IDF", Journal of Physics, Vol. 1486, No. 7, PP. 1-7, 2020.

[27] http://www.tfidf.com/

[28] Aytug Onan, and Serdar Korukoglu, "A Feature Selection Model based on Genetic Rank Aggregation for Text Sentiment Classification", Vol. 43, No. 1, PP. 25-38, 2017.

[29] Cláudia Pascoala, M. Rosário Oliveiraa, António Pachecoa, and Rui Valadas, "Theoretical Evaluation of Feature Selection Methods Based on Mutual Information", Neurocomputing, Vol. 226, PP. 168–181, 2017.

[30] Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E. Alsaadi, "Evaluation of Feature Selection Methods for Text Classification with Small Datasets Using Multiple Criteria Decision-Making Methods", Applied Soft Computing Journal, Vol. 86, PP. 1-14, 2020.

[31] S. Paudel, P. W. C. Prasad, Abeer Alsadoon, MD. Rafiqul Islam, and Amr Elchouemi, "Feature Selection Approach for Twitter Sentiment Analysis and Text Classification Based on Chi-Square and Naïve Bayes", International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI, Switzerland, PP. 281-298, 2018.

[32] Mandieh Labani, Parham Moradi, Fardin Ahmadiz, and Mandi Jalili, "A Novel Multivariate Filter-Method for Feature Selection in Text Classification Problems", The Journal of Engineering Application of Artificial Intelligence, Vol. 70, PP. 25-37, 2018.

[33] https://scikit-learn.org/stable/modules/generated/ sklearn. feature selection.fregression.html

[34] https://github.com/hadyelsahar/large-arabic-sentiment-analysis - resouces/tree/master/datasets

[35] M. Baygin, "Classification of Text Documents Based on Naïve Bayes Using N-gram Features", International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, DOI:10.1109/idap.2018.8620853, 2018.

[36] M. Asghar, A. Khan, S. Ahmed, M. Qasim, and I. Khan, "Lexicon-Enhanced Sentiment Analysis Framework Using Rule-Based Classification Scheme", Journal PLOS ONE, PP. 1-22, February 23, 2017, https://journals.plos.org/ plosone/article?id=10.1371/journal. pone.0171649.

[37] Kangfeng Zheng, and Xiujuan Wang, "Feature Selection Method with Joint Maximal Information Entropy between Features and Classes", Pattern Recognition, Vol. 77, PP. 20-29, 2018.

تصنيف آراء المستخدمين وتعليقاتهم على فيسبوك باستخدام أساليب تعلم الآلة

إبراهيم روبى
معهد بحوث الإلكترونيات
القاهره – جمهورية مصر العربية
ibrahimelgazwy@gmail.com

محمد نور
معهد بحوث الإلكترونيات
القاهره – جمهورية مصر العربية
mnour99@hotmail.com

محمد بدوى
كلية الهندسة الإلكترونية- جامعة المنوفية
المنوفية – جمهورية مصر العربية
mohamed.badawi@el-eng.menofia.edu.eg

إحسان عابد
معهد بحوث الإلكترونيات
القاهرة – جمهورية مصر العربية
eabed03@yahoo.com

الملخص العربى

يقدم هذا العمل تحليلا لأربعة أنواع من مصنفات تعلم الآلة الموجة وهى SVM NB, SGD, LR، والتى تم تطبيقها على عينة اختبارية من البيانات تحوى على ما يزيد عن ثمانية آلاف تعليقا ورأيا مكتوبة بالعربية من قبل مستخدمى فيسبوك للتعبير عن آرائهم فى خدمة المطاعم. كما تبنى العمل أيضا تحليلا لأربع طرق مختلفة لاستخلاص الصفات المعبرة عن تلك الآراء والتعليقات المكتوبة، لما لتلك الصفات من تأثير كبير على أداء مصنفات التعلم، وهذه الطرق تعتمد على: وزن العناصر وأهميتها، درجة الارتباط، مقياس χ^2 ، ودرجة التأثير والتداخل المعلوماتى. وقد تم استخدام بعض المعايير مثل: درجة الدقة، درجة الاسترجاع، مقياس F، وزمن التعلم، لتقييم أداء مصنفات التعلم، وكذا طرق استخلاص الصفات. ومن خلال التجارب العملية والمقارنه، وجد أن أداء مصنف SVM كان هو الأفضل من المصنفات الأخرى، كما أن طريقة استخلاص الصفات المعتمدة على درجة الارتباط بين صفات الآراء ونوع التصنيف كانت هى الأفضل من الطرق الأخرى، كما أن مصنف SVM وطريقة استخلاص الصفات المعتمدة على درجة الارتباط قد استهلكا وقتا أكبر فى مرحلة التعلم مقارنة بالأساليب الأخرى. ومن المتوقع التوصل لنفس التوصيات عند تطبيق نفس الأساليب ونفس طرق استخلاص الصفات على عينات اختبارية آخرى.

الكلمات الدالة: تعلم الآلة الموجة، أساليب التصنيف، طرق استخلاص الصفات، تعليقات وآراء مستخدمى فيسبوك، تقييم الأداء