Journal of Theoretical and Applied Information Technology
20th January 2015. Vol.71 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# HYBRID STATISTICAL RULE-BASED CLASSIFIER FOR ARABIC TEXT MINING

[1] ABDULLAH S. GHAREB, [2]ABDUL RAZAK HAMDAN, [3]AZURALIZA ABU BAKAR, [4]MOHD RIDZWAN YAAKUB

[1, 2, 3,4] Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia,
43600, UKM, Bangi, Selangor, Malaysia
E-mail: [1]aghurieb@yahoo.com , [2]arh@ukm.edu.my, [3]azuraliza@ukm.edu.my,
[4]ridzwanyaakub@ukm.edu.my

**ABSTRACT**

Text categorization is one of key technology for organizing digital dataset. The Naiv Bayes (NB) is popular categorization method due its efficiency and less time complexity, and the Associative Classification (AC) approach has the capability to produces classifier rival to those learned by traditional categorization techniques. However, the independence assumption for text features and the omission of feature frequencies in NB method violates its performance when the selected features are not highly correlated to text categories. Likewise, the lack of useful discovery and usage of categorization rules is the major problem of AC and its performance is declined with large set of rules. This paper proposed a hybrid categorization method for Arabic text mining that combines the merits of statistical classifier (NB) and rule based classifier (AC) in one framework and tried to overcome their limitations. In the first stage, the useful categorization rules are discovered using AC approach and ensure that associated features are highly correlated to their categories. In the second stage, the NB is utilized at the back end of discovery process and takes the discovered rules, concatenates the associated features for each category and classifies texts based on the statistical information of associated features. The proposed method was evaluated on three Arabic text datasets with multiple categories with and without feature selection methods. The experimental results showed that the hybrid method outperforms AC individually with/without feature selection methods and it is better than NB in few cases only with some feature selection methods when the selected feature subset was small.

**Keywords:** *Text Categorization, Naïve Bayes, Associative Classification, Hybrid Categorization Method, Feature Selection.*

## 1. INTRODUCTION

Data mining is an important process for extracting the useful knowledge in order to help users to manage, analyze, retrieve, and control the flow of data and information. The categorization is a data mining task that can help for natural language processing and mining of useful knowledge for several application domains. It has been applied successfully in several domains; for example speech domain such as speaker state classification [1] and speech recognizers with topic categorization [2], biomedical dataset categorization [3] and image categorization [4]. However, most of the information today is stored as texts [5]; thus, text mining is an important topic of research for mining knowledge from textual dataset. The text categorization or the process of automatically predicting the valid categories of texts has become one of the major techniques for organizing and managing Arabic information [6]. The traditional categorization techniques are the most frequently used for text categorization such as Naïve Bayes (NB) [7] [8] [9], Support Vector Machine (SVM) [6] [10] [11], Neural Network (NN) [12] and K Nearest Neighbor (KNN) [13] [14] [15].

Among these categorization methods, the NB is one of the successful methods and it is widely used for categorization due to its efficient computation and good performance [7]. Despite its advantages, NB assumes that features are independent and it only needs the variances of features for text

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

categories to make prediction, also, the frequent features are not considered by NB, however, the feature frequencies can reflect the actual contribution of features in a given text documents [16]. This issue can violate the NB performance specially when the features are highly associated, and so, it is sensitive to feature selection (FS) process [7] [17].

On the other hand, the rule based categorization techniques are rarely investigated for Arabic text categorization [18]. The effort in this type is very limited when comparing to other traditional techniques but they achieve competitive results [18] [19] [20]. Nevertheless, the insufficient and inaccurate categorization rules have negative effect on the categorization performance [18] [19] [20] [21]. The Associative classification (AC) approach depends on the association rules to categorize text dataset; it has been used for Arabic text categorization in few works [19] [20] [22]. The AC is a good method to discover the frequent features and it is able to reduce the high dimensionality of features through the association rule mining concept. On the other hand, the previous researches point that the AC performance is declined due to large set of categorization rules; in addition, they ignore some useful information that can help for prediction. Moreover, when adopting AC, many challenges should be considered such as the discovery process of high confidence rules which require long time, the producing of useful rules that comprise the most portions of the studied text categories and the utilizing of suitable forecasting method [19] [22] [23]. Consequently, the lack of useful categorization rules that cover text categories with reasonable set of rules and the prediction method that used by AC should be enhanced when rule approach is utilized for Arabic text categorization.

Usually, the combination of the merits of several categorization techniques can achieve better performance by incorporating both techniques in one framework. The hybrid mechanism is introduced as a direction of research to improve the accuracy and efficiency of an individual categorization technique [17]. Several hybrid methods are proposed, among them, Isa et al. (2009) [24] showed that the combining of NB and self organizing map superior to the basic NB. Another study [25] also proven that the hybrid NB-SVM is significantly better than NB individually. The authors in [26] introduced a set of KNN classifiers for category prediction and then the test documents are assigned to multi-categories based on a Bayesian voting mechanism. In addition, Miao et al. (2009) [27] combined the Rocchio classifier and KNN to improve the categorization performance. Likewise, Li, Miao & Wang (2011) [28] combined SVM with KNN to construct a hierarchical algorithm for text categorization. Further information for hybrid classifier and combination of the categorization techniques can be found in literature [17] [28] [29].

Feature Selection (FS) is an essential phase in text categorization due the high dimensionality and the existing of misleading features that degrade the categorization performance. The FS methods are either filter or wrapper; however, the filter approach is frequently employed with text categorization due its reasonable computational cost. Several filtering methods are proposed and used to select the most informative features for text categorization, for example, Class Discriminating Measure (CDM) [7] , F-measure (FM) for features inside training dataset [6][30], Galavotti, Sebastiani and Simi measure (GSS) [31], Odd Ratio (OR) [32] [33], discriminative power and GINI index based term frequency measures [16], term frequency and inverse document frequency (TF-IDF) [34] [35], comprehensive measure for feature selection [36], and many more.

In this paper, we focus on two problems of two different categorization methods; the independent assumption for NB which affects its performance if the correlated features are not selected and the lack of useful and accurate categorization rules in the AC situation. These problems can be avoided by improving rule discovery and prediction processes. Therefore, we introduce a hybrid method named hybrid ACNB that merges correlation or association between features with probabilistic information for features. The features are extracted based on association between them using AC and used by NB which assumes that features are independent. Furthermore, we take into account the high dimensionality problem and use the FS methods to reduce text dimensionality. The paper presents comparison between the proposed hybrid method with AC and NB individually and examine the effect of five FS methods (CDM, FM, GSS, OR and TF-IDF) on the categorization performance for various dimensions of features.

Journal of Theoretical and Applied Information Technology
20th January 2015. Vol.71 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The rest of this paper is organized as follows: Section 2 describes the proposed hybrid categorization method. Section 3 discusses the experimental results and Section 4 presents the conclusion of this paper.

## 2. HYBRID ACNB CATEGORIZATION METHOD

The hybridization process of two categorization techniques aims to incorporate their advantages together and initiate new hybrid categorization method. Two methods can be used to introduce a hybrid method, either use one method for learning and the other method for revising and predicting or combining the functions of both methods in a single hybrid function. In this paper, we adopt the first choice, a hybrid categorization method based on the combination of AC and NB classifier is developed. The developing process starts by discovering the associative features of Arabic text datasets, where for each category a set of features is discovered by the association rule mining algorithm and then revised. Then the discovered association rules for all categories are used by NB to discriminate the different categories of text datasets. In the following subsections we describe NB, AC and the hybrid ACNB categorization methods.

### 2.1 Naïve Bayes

The multinomial NB usually outperforms other variations of NB [37]; therefore, we use it at the front end of AC to predict the valid categories of text documents. For a given document $d$; in text categorization, the aim is to find the correct category of $d$. The best category $c$ of $d$ in NB categorization method is the category that has the highest probability that indicates the document $d$ belongs to category $c$ and it is calculated as follows:

$$c = max_{c \in C} P(c \mid d) = max_{c \in C} P(c) \Pi_i P(f_i \mid c) \quad (1)$$

$$P(f_i \mid c) = (Fc + cons) / (Nc + Fv) \quad (2)$$

where $P(c|d)$ is the probability that a given document $d$ belongs to category $c$. $P(c)$ is the probability of a category $c$ which computed as the ratio of documents that belongs to category $c$ among the total number of documents. $P(f_i|c)$ is the probability of a set of features $f_i$ given a category $c$. $Fc$ is the number of times the feature $i$ occurs in category $c$, $Nc$ is The number of features in category $c$, $Fv$ is the number of features in a given vocabulary and $cons$ is the positive constant, set to one to avoid zero probability.

### 2.2 Associative Classification

In this paper, the Apriori based algorithm is applied on text document that represented by a set of features from the previous phase (text preprocessing and FS phase). This algorithm performs an iterative search over a collection of text document to discover the frequent features that distinguish each category from others. The discovery process of candidate and frequent m-feature set is performed until no further frequent feature set can be generated. The rules that form AC classifier are only rules that indicate category label; $f_m \rightarrow c_i$, where $f_m$ is a set of frequent features ($f_1$& $f_2$ &…& $f_m$) that represent documents in the training dataset (rule body), and $c_i$ is the category that associated with these features (rule head). Each rule must have support and confidence, the rule support is defined as the percentage of documents that contain features of each rule, whereas, the rule confidence reflects the degree of confidence that this rule belongs to category $c_i$. In the categorization phase, the majority voting method is utilized for categorization; in this method, all rules that cover test document are retained, and if all retained rules associate with only one category, this category will assigned to test document. However; when the retained rules are distributed among two or more categories, then these rules divided into n sub-sets by categories, and the test document is assigned to the category which has the largest number of rules. Details for the AC algorithms and its steps can be found in literature [19] [22] [38].

### 2.3 Proposed Hybrid ACNB Categorization Method

The AC is good technique to discover association between text features and to reduce the huge amount of unneeded features [42]. The NB is efficient in term of computational cost and has good performance. The combination of AC and NB classifier is proposed to merge their strengths and improve prediction power of the categorization rules. Figure 1 depicts the architecture of the proposed hybrid ACNB method. In the hybrid method ACNB, the AC approach [19] [22] is used to discover the categorization rules from training datasets, the extracted rules are composed from a set of features after text preprocessing, the next step is to compare the discovered association rules against the original features and the matched features in each rule are returned with its prior probabilistic information such as their frequency and weight, this information will be used later by NB classifier for decision making. For each category of the studied text, a validity process is applied that ensure each category has enough set of

Journal of Theoretical and Applied Information Technology
20th January 2015. Vol.71 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

categorization rules that have strong confidence. The effectiveness of the discovered rules is initially tested by the number of training samples that are covered by the categorization rules, and the useless rules and interlaced with more than two categories are rejected. Finally, the NB uses the revised associated features to classify the test documents to its proper category.

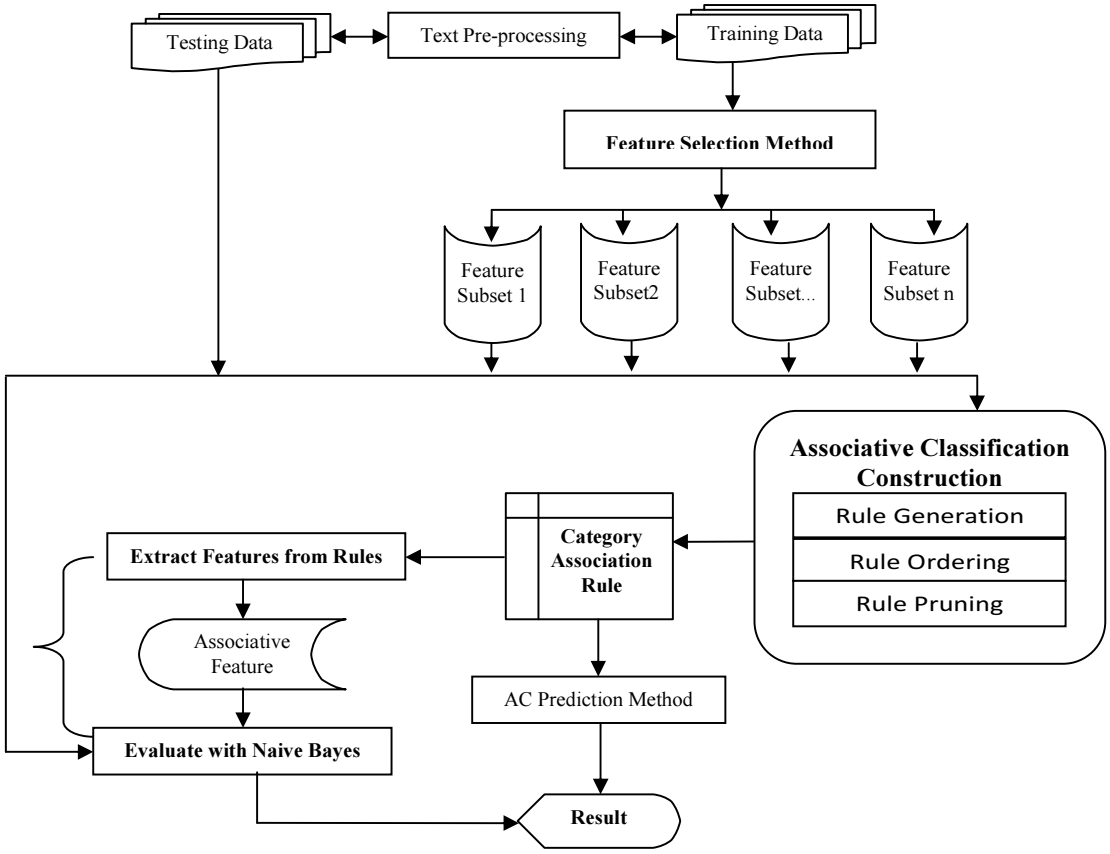



*Figure 1: Architecture of the Hybrid ACNB Categorization Method*

The steps of hybrid ACNB algorithm are summarized as follow:

- **Input**: a collection of text documents D for all categories;
- **Output**: a set of categorization rules for all categories, a set of associated features and decision (test documents with their categories).
- **Step1**: Text preprocessing:

For each category $c_i$ in text dataset:
For each document in category $c_i$ :

**a.** Apply normalization process; Split the text into a token (word); and extract the Arabic nouns as a candidate features [22].

**b.** Apply stop words removal process and stemming algorithm to ensure that text is clean from function words and to reduce text sparsity [19].

**c.** Represent text documents as vectors of document-terms to corresponding category.

- **Step 2**: Feature Selection:

**a.** Measures the quality of features based on the following five FS methods that compute rank for each feature in dataset:

$$CDM = |\log A/B| \quad (3)$$

$$FM = 2AB/A+B \quad (4)$$

$$GSS = ((A * D) - (B * C)) \quad (5)$$

$$OR = (A * (1-B)) / (B * (1-A)) \quad (6)$$

$$TF\text{-}IDF = TF(f, c_i) * \log (N/(A+B)) \quad (7)$$

where $N$ is the total number of training documents in the collection, $A$ is the probability of feature $f$ given category $c_i$, $B$ is the probability of feature $f$ in other categories excluding $c_i$, $C$ is the probability of the absent of feature $f$ given category $c_i$, $D$ is the

Journal of Theoretical and Applied Information Technology
20th January 2015. Vol.71 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

probability of the absence of feature *f* in other categories excluding $c_i$ and *TF (f, $c_i$)* is the frequency of feature *f* given category $c_i$.

**b.** Order all features in decreasing order based on their importance.

**c.** Select the top ranked features (a given number of features).

**d.** Return the feature vectors for all categories and save each document with its returned features.

- **Step 3**: Categorization Rules Generation:
  Generate the categorization rules and extract the associated features (choose either with full features or with a subset of features) [22]:
  **a.** Model each document as a transaction and its features as transaction items
  **b.** Perform an iterative search through documents *D* and generate the category association rule using Apriori based algorithm.
  **c.** Order the category association rule according to its support, confidence, and length.
  **d.** Remove the redundant rules, the useless rules and rules that associated with more than two categories.
  **e.** Group the revised rules according to text categories and ensure that rules cover all training categories.
- **Step 4**: New Document Categorization:
  **a.** Concatenate the associated features for each category in a one list and calculate their probabilistic information.
  **b.** Use NB to classify the new documents to their valid categories based on the statistical information of the associated features.

In the hybrid ACNB method; the learning process and the categorizing of new document are highly depending on the associated feature sets that generated by AC method. Therefore, the method efficiency is depending on the predictive power of the extracted categorization rules and their probabilistic information.

## 3. EXPERIMENTAL RESULTS

### 3.1 Datasets and Evolution Measure

To evaluate the proposed hybrid ACNB method, the experiments are conducted on three Arabic text datasets. The datasets include two relatively small datasets ("Al-Jazeerah and "Akhbar Al-Khaleej text datasets") with five and four categories, respectively, and one large dataset "Al-waten text dataset" with five categories [39] [40]. Table 1 presents the distribution of documents to different categories with each dataset. The texts in each dataset are processed according to Arabic text preprocessing tools and modeled as vectors of documents-terms to category.

The result evaluation is carried out in terms of categorization performance (macro-average values). The categorization effectiveness in terms of precision ($P_i$), recall ($R_i$) and F-measure ($F_i$) for each category is computed as follow:

$$P_i = A/D \quad (8)$$

$$R_i = A/T \quad (9)$$

$$F_i = (2\ P_i\ R_i) / (P_i + R_i) \quad (10)$$

where *A* is the number of documents correctly assigned to category $c_i$, *D* is the number of documents assigned to category $c_i$ and *T* is the total documents in category $c_i$.

The macro-average precision ($P_{macro}$), macro-average recall ($R_{macro}$) and macro-average F-measure ($F\text{-}m_{macro}$) are employed in this paper to evaluate the performance across multiple categories C. They are identified as in the following equations:

$$P_{macro} = \sum_{i=1}^{C} Pi\ /C \quad (11)$$

$$R_{macro} = \sum_{i=1}^{C} Ri\ /C \quad (12)$$

$$F\text{-}m_{macro} = \sum_{i=1}^{C} Fi\ /C \quad (13)$$

*Table 1: Distribution of "Alwatan", "Al-jazeera" and "Akhbar-Alkhaleej" Datasets*

| "Alwatan" text dataset | | | "Al-jazeera" text dataset | | "Akhbar-Alkhaleej" text dataset | |
|---|---|---|---|---|---|---|
| **Category** | **# documents** | **# words** | **Category** | **# documents** | **Category** | **# documents** |
| Religion | 3860 | 3,144,828 | Economy | 300 | Economy | 273 |
| Economy | 3468 | 1,482,009 | Politics | 300 | Sports | 429 |
| Sports | 4550 | 1,447,889 | Sport | 300 | Local news | 720 |
| Culture | 2782 | 1,411,218 | Science | 300 | Int. news | 286 |
| Int. news | 2035 | 865,671 | Art | 300 | *Total* | 1708 |
| *Total* | 16695 | 8,351,615 | *Total* | 1500 | *# words* | 746,307 |
| | | | *# words* | 389,766 | | |

Journal of Theoretical and Applied Information Technology
20th January 2015. Vol.71 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

### 3.2 Results and Discussion

#### 3.2.1 Result of hybrid ACNB without feature selection

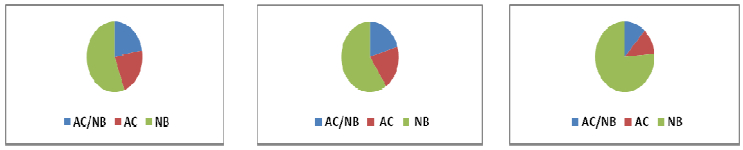
This section presents the results of hybrid ACNB categorization method and it also shows a comparison between the hybrid ACNB with AC and NB individually. The minimum support and minimum confidence are set to 10% and 50% respectively. Table 2 presents the results of hybrid ACNB, NB, and AC for all features after text preprocessing (without any FS method), note that Al-Jazeera dataset, Akhbar Al-Khaleej dataset and Al-waten dataset are abbreviated D1, D2 and D3, respectively. The annotations $P_{macro}$, $R_{macro}$ and $F\text{-}m_{macro}$ mean macro-average precision, macro-average recall and macro-average F-measure, respectively. As shown in Table 2, for different datasets, the hybrid ACNB achieves its best performance with D1 in terms of all measures and the lowest performance with D2. The reason is that Al-Jazeera text datasets include more associated features that have the ability to distinct the different categories of texts in this datasets. The performance of hybrid ACNB for D3 is better than when it is with D2 but its lower than it for D1. This indicates that text characteristics have real impact on the categorization performance.

In addition, the comparison results between the hybrid ACNB with AC and NB showed that Hybrid ACNB acquired better performance than AC in terms of all performance measures (P macro, R macro and F-m macro) but its performance is lower than NB with all studied text datasets. The behavior of Hybrid ACNB is composed from AC and NB behaviors, it take AC advantages for training and feature extraction/reduction and it employ NB for testing. The NB tests the usefulness of the discovered rules based on Bayesian categorization theorem. The performance of hybrid categorization method should give better results than the combined techniques individually, but this suggestion is not always true, the comparison results emphasized this observation. However, the results of them may be different with FS methods because NB is more sensitive to FS. The results indicate that NB alone can work well with a large subset of features because the effect of independent assumption about features is reduced with full features. Despite this point, the hybrid ACNB has advantages over NB in term of final categorization (decision), it just use a subset of associated features with all required information about features and this subset can be extended by adding more training documents, and so, it can speed up the prediction process. However, in the NB case with high dimensional dataset, the use of all features is impractical because it has a negative effect on the computer resource, it will need a large memory space and high computations cost. To illustrate this point, Figure 2 shows the classifier size (feature vectors) that used by each categorization methods for each datasets.

It clears from Figure 2 (a-c) that the required space at the final stage of categorization process for hybrid ACNB and AC is smaller than that for NB. The reason is that the feature space is reduced after training by AC by about 60.23%, 65.90%, and 84.54% for D1, D2, and D3, respectively. In addition, the categorization rules are also contain reduced subset of features, the maximum length of rule is less than ten features and most of rules have small sizes, ranged from 2 to 6 features, but in case of NB individually, the full features are participating for categorization without any reduction except that performed by preprocessing tools with all categorization methods.

*Table 2: Performance (Macro Average %) of Hybrid ACNB, NB and AC Classifiers with All Features for each Dataset*

| Data set | Hybrid AC/NB | | | NB | | | AC | | |
|---|---|---|---|---|---|---|---|---|---|
| | $P_{macro}$ | $R_{macro}$ | $F\text{-}m_{macro}$ | $P_{macro}$ | $R_{macro}$ | $F\text{-}m_{macro}$ | $P_{macro}$ | $R_{macro}$ | $F\text{-}m_{macro}$ |
| D1 | 85.906 | 85.00 | **85.451** | 88.101 | 87.333 | **87.715** | 71.95 | 61.08 | 65.824 |
| D2 | 73.542 | 79.894 | 76.586 | 81.439 | 86.784 | 84.029 | 61.489 | 62.621 | 62.048 |
| D3 | 83.339 | 83.888 | 83.613 | 88.084 | 87.330 | 87.706 | 73.631 | 64.681 | **68.748** |




(a) D1 (b) D2 (c) D3

*Figure 2: Classifier Size of each Categorization Method for each Dataset*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

#### 3.2.2 Effect of FS methods on the hybrid ACNB performance

The proposed method is tested with FS methods to show if the AC is good enough to reduce text dimensionality in case of hybrid ACNB and AC and to show their effect on the performance of hybrid ACNB, NB and AC. The FS methods include CDM, FM, GSS, OR and TF-IDF, five feature dimensions (1000 to 5000) are selected with each FS method.

The results in Table 3 to Table 5 show the performance of all categorization methods based on five FS methods for three text datasets (D1, D2 and D3). From results, in case of all datasets, it can be seen that the hybrid ACNB work better with TF-IDF and FM with slight difference between them, they achieved the best average of macro F-measure cross different feature subsets. The third highest average performance of hybrid ACNB is acquired by GSS in case of D1 and D3, and by OR in case of D2.

An important point that should be noted is that in case of D1 and D3 (Table 4-Table 5) no one of the used FS methods outperformed the hybrid ACNB without feature selection, the performance of hybrid ACNB is not improved by utilizing FS methods with the varying feature subsets. The behind reason is that the categorization rules that form the associative features is more precise with full features, and so, the pattern of text categories is clearly appeared and the features are strongly correlated to their categories, however, when use a subset of feature to participate for rule generation; the number of associative features is decreased which affect the categorization decision.

This indicates that the AC approach in the first phase is good enough in case of D1 and D3 to produces the relevant associative features. Nevertheless, in case of D2 (Table 6), the performance of hybrid ACNB is slightly improved in term of average across different subsets with all FS methods except GSS. Also, its performance with CDM and OR for 4000 and 5000 features is significantly better than without FS. This happened because the extracted features from categorization rules with full features have small support from training documents and they are not enough to make accurate categorization, but the result is improved when the ACNB is guided by some FS methods, so the rule support and confidence are increased and the extracted features are more related to their categories. Generally, the computational cost can be reduced by using the FS methods because a subset of features is only participated for classifier construction.

In case of AC alone, the TF-IDF achieved the best average results with all datasets, the other FS methods have a negative effect on the AC performance and their results are always lower than AC without FS methods. This emphasizes that AC approach is enough alone to produces the relevant features.

*Table 3: Performance (macro average F-measure %) of Hybrid AC/NB, AC and NB with Five FS Methods for D1 Dataset*

| Classifier | FS Method | #Feature 1000 | 2000 | 3000 | 4000 | 5000 | Average |
|---|---|---|---|---|---|---|---|
| Hybrid | CDM | 67.354 | 69.885 | 74.308 | 69.184 | 71.495 | 70.4452 |
| ACNB | FM | 82.942 | 83.153 | 83.192 | 82. 838 | 82.838 | 83.0313 |
| | GSS | 81.964 | 81.279 | 80.899 | 81.597 | 81.289 | 81.4056 |
| | OR | 71.462 | 70.121 | 72.180 | 71.079 | 71.143 | 71.197 |
| | TF-IDF | 83.634 | 83.183 | 83.236 | 83.194 | 83.534 | 83.3562 |
| AC | CDM | 40.00 | 38.376 | 67.982 | 60.181 | 56.346 | 52.577 |
| | FM | 55.823 | 55.823 | 55.823 | 55.823 | 55.823 | 55.823 |
| | GSS | 59.803 | 58.873 | 58.743 | 58.649 | 58.649 | 58.9434 |
| | OR | 40.00 | 38.223 | 52.594 | 49.085 | 57.556 | 47.4916 |
| | TF-IDF | 73.332 | 77.756 | 75.660 | 77.328 | 77.480 | 76.3112 |
| NB | CDM | 66.143 | 69.416 | 75.392 | 78.917 | 83.936 | 74.7608 |
| | FM | 87.678 | 88.535 | 88.551 | 88.207 | 88.207 | 88.2356 |
| | GSS | 87.754 | 88.523 | 87.557 | 87.894 | 87.894 | 87.9244 |
| | OR | 67.993 | 69.228 | 74.236 | 78.744 | 82.988 | 74.6378 |
| | TF-IDF | 87.625 | 89.921 | 89.869 | 88.207 | 88.207 | 88.7658 |

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 4: Performance (macro average F-measure %) of Hybrid AC/NB, AC and NB with Five FS Methods for D2 Dataset*

| Classifier | FS Method | #Feature 1000 | 2000 | 3000 | 4000 | 5000 | Average |
|---|---|---|---|---|---|---|---|
| Hybrid | CDM | 60.110 | 60.111 | 76.577 | 91.916 | 94.083 | 76.5594 |
| ACNB | FM | 77.359 | 76.949 | 77.002 | 77.341 | 77.341 | 77.1984 |
| | GSS | 75.161 | 75.068 | 74.817 | 73.745 | 75.377 | 74.8336 |
| | OR | 60.110 | 60.111 | 76.577 | 92.462 | 94.460 | 76.744 |
| | TF-IDF | 78.394 | 77.441 | 77.064 | 76.938 | 76.688 | 77.305 |
| AC | CDM | - | 46.260 | 67.298 | 68.396 | 66.960 | 49.7828 |
| | FM | 42.109 | 42.048 | 42.048 | 42.048 | 42.048 | 42.0602 |
| | GSS | 45.669 | 44.049 | 43.497 | 43.497 | 43.447 | 44.0318 |
| | OR | 25.00 | 62.821 | 64.913 | 57.303 | 56.522 | 53.3118 |
| | TF-IDF | 60.766 | 60.444 | 66.426 | 66.799 | 67.599 | 64.4068 |
| NB | CDM | 48.712 | 83.607 | 77.634 | 73.925 | 74.805 | 71.7366 |
| | FM | 84.218 | 87.054 | 86.806 | 86.949 | 87.976 | 86.6006 |
| | GSS | 82.435 | 84.140 | 84.838 | 82.709 | 82.497 | 83.3238 |
| | OR | 48.712 | 83.001 | 74.927 | 72.088 | 73.161 | 70.3778 |
| | TF-IDF | 83.341 | 87.303 | 86.622 | 87.162 | 88.024 | 86.4904 |

*Table 5: Performance (macro average F-measure %) of Hybrid AC/NB, AC and NB with Five FS Methods for D3 Dataset*

| Classifier | FS Method | #Feature 1000 | 2000 | 3000 | 4000 | 5000 | Average |
|---|---|---|---|---|---|---|---|
| Hybrid | CDM | 63.327 | 71.702 | 71.906 | 74.733 | 74.845 | 71.3026 |
| ACNB | FM | 79.551 | 79.595 | 79.677 | 79.652 | 79.644 | 79.6238 |
| | GSS | 77.657 | 77.645 | 77.832 | 77.889 | 78.067 | 77.818 |
| | OR | 65.425 | 72.801 | 76.350 | 76.716 | 76.728 | 73.604 |
| | TF-IDF | 79.335 | 79.308 | 79.532 | 79.576 | 79.646 | 79.4794 |
| AC | CDM | - | 19.967 | 49.836 | 64.184 | 62.776 | 39.3526 |
| | FM | 48.748 | 48.748 | 48.748 | 48.748 | 48.748 | 48.748 |
| | GSS | 51.714 | 51.539 | 50.723 | 50.649 | 50.642 | 51.0534 |
| | OR | - | 37.938 | 58.034 | 62.484 | 60.241 | 43.7394 |
| | TF-IDF | 63.006 | 67.226 | 67.959 | 72.183 | 71.821 | 68.439 |
| NB | CDM | 38.818 | 65.642 | 75.789 | 75.952 | 73.454 | 65.931 |
| | FM | 88.969 | 88.602 | 88.878 | 87.899 | 87.784 | 88.4264 |
| | GSS | 87.838 | 88.854 | 88.077 | 87.318 | 87.107 | 87.8388 |
| | OR | 38.818 | 65.949 | 77.415 | 75.832 | 73.682 | 66.3392 |
| | TF-IDF | 88.849 | 88.477 | 89.882 | 88.991 | 88.771 | 88.994 |

In case of NB, it get the best results with FM, TF-IDF and GSS for all datasets and their results are better than NB without FS excluding GSS in the case of D2, in which its average result is slightly lower than NB with all features.

As the results indicate, the hybrid ACNB is always outperformed AC individually with all FS methods for all datasets. In addition, the hybrid ACNB with CDM and OR has advantage over NB specially for small feature subset, for example, the performance of hybrid ACNB is better than NB

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with 1000 features for all datasets and also their average results with ACNB are significantly better than that with NB. Furthermore, in case of D2, the hybrid ACNB with these methods (i.e. CDM and OR) achieved better performance than NB when the feature subsets are large (4000 and 5000). The AC in some situations has poor performance for three reasons, (i) no rules are discovered specially with small feature subsets, (ii) the discovered rules does not cover all categories, and (iii) the categorization decision is confused with large number of categorization rules. In the proposed hybrid ACNB these drawbacks are minimized by ensuring that the discovered rules covering all text categories and the associated features are highly correlated to their target categories, also we employed statistical method for prediction which not confused by the number of rules.

Based on the experimental results, it can be concluded that the utilizing of hybrid ACNB can create new categorization method that incorporate the features association with their probabilistic information and achieve promising results. This method merges correlation or association between features with feature independency in which the features are extracted based on association between them and used by NB which assume that features are independent. The AC reduces the large number of features that considered for text categorization and returns only features that have strong association to be used for categorization. The proposed method ensures that all categories are covered by the discovered categorization rules, and so, the possibility of fail categorization or low performance is reduced. The proposed method achieves promising result compared to AC and NB individually. The ACNB was always better than AC, but it is better than NB in few cases and comparable in some cases. In addition, the experimental results confirmed that each categorization method had inconsistent results for different text datasets and their performances were affected by text characteristics. Furthermore, when comparing the different FS methods for different dimensions of features, their results were unstable for different categorization methods, different datasets and different dimensions. Overall, in text mining field, it should consider the text properties, the suitability of the used FS methods and the categorization algorithm behavior.

## 4. CONCLUSION

This paper presented hybrid categorization method based on the combination of AC and NB categorization methods. In the first stage of hybrid ACNB, the AC is employed to generate the useful categorization rules; the rules are revised to ensure they are complete and highly associated with their categories. In the second stage, the NB classifier is utilized at the back end of rule discovery process and extracts the best associated features for each category, then; it classifies texts based on the statistical information of the extracted associated features. The proposed method is examined with full features and with five feature selection methods under different dimensions of features. The experimental results showed that the hybrid ACNB method outperformed AC individually for all cases, however, it superior NB in few cases only with regard to categorization performance.

## REFERENCES:

[1] D.Y. Huang, Z. Zhang, and S.S. Ge, "Speaker state classification based on fusion of asymmetric simple partial least squares (SIMPLS) and support vector machines", *Computer Speech and Language,* Vol. 28, No. 2, 2014, pp. 392-419.

[2] M-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an HMM-based self-organizing unit recognizer with applications to topic classification and keyword discovery", Computer Speech and Language, Vol. 28, No. 1, 2014, pp. 210-223.

[3] L. Shi, L. Xic, X. Ma, M. Weng, and X. Hu, "A novel ensemble algorithm for biomedical classification based on Ant Colony Optimization", Applied Soft Computing, Vol. 11, 2011, pp. 5674–5683.

[4] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation", *In Computer vision and pattern recognition*, *CVPR* 2008, IEEE, pp. 1-8.

[5] V. Korde, and C.N. Mahender, "Text classification and classifiers: a survey", *International Journal of Artificial Intelligence and Applications,* Vol. 3, No. 2, 2012, pp. 85-99.

[6] A. Mesleh, "Feature sub-set selection metrics for Arabic text classification", *Pattern Recognition Letters*, Vol. 32, No. 14, 2011, pp. 1922-1929.

[7] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes", *Expert Systems with Applications*, Vol. 36, No. 3, 2009, pp. 5432-5435.

[8] A.M. Hattab, and A.K. Hussein, "Arabic Content Classification System using statistical Bayes classifier with Words Detection and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Correction", *World of Computer Science and Information Technology Journal,* Vol. 2, No. 6, 2012, pp. 193-196.

[9] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve Bayesian based on Chi Square to categorize Arabic data", *In: Proceedings of the 11th International Business Information Management Association Conference (IBIMA)* Vol. 10, 2009, pp. 158-163.

[10] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", *In: Proceedings of ECML'98: 10th European Conference on Machine Learning*, Springer, 1998, pp. 137–142.

[11] V. Mitra, C.J. Wang, and S. Banerjee, "Text classification: A least square support vector machine approach", *Applied Soft Computing*, Vol. 7, No. 3, 2007, pp. 908-914.

[12] F. Harrag, and E. Al-Qawasmah, "Improving Arabic Text Categorization using Neural Network with SVD", *Journal of Digital Information Management (JDIM),* Vol. 8, No. 2, 2010, pp. 125-135.

[13] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved K nearest-neighbor algorithm for text categorization", *Expert Systems with Applications,* Vol. 39, No. 1, 2012, pp. 1503-1509.

[14] M.M. Abu Tair, and R.S. Baraka, "Design and Evaluation of a Parallel Classifier for Large-Scale Arabic Text", *International Journal of Computer Applications,* Vol. 75, No. 3, 2013, pp. 13-20.

[15] K. Abidi, Z. Elberrichi, and Y.G. Tlili, "Arabic Text Categorization: A Comparative Study of Different Representation Modes", *Journal of Theoretical and Applied Information Technology*, Vol. 38, No. 1, 2012, pp: 1–5.

[16] N. Azam, and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization", *Expert Systems with Applications*, Vol. 39, 2012, pp. 4760–4768.

[17] B. Baharudin, L.H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification", *Journal of advances in information technology,* Vol. 1, No. 1, 2010, pp. 4-20.

[18] M. Al-diabat, "Arabic Text Categorization using Classification Rule Mining", *Applied Mathematical Sciences,* Vol. 6, No. 81, 2012, pp. 4033-4046.

[19] Q.A. Al-Radaideh, E.M. Al-Shawakfa, A.S. Ghareb, and H. Abu-Salem, "An Approach for Arabic Text Categorization using Association Rule Mining", *International Journal of Computer Processing of Languages,* Vol. 23, No. 01, 2011, pp. 81-106.

[20] S. Al-Saleem, "Associative classification to categorize Arabic data sets", *The International Journal of ACM Jordan,* Vol. 1, 2010, pp. 118-127.

[21] F. Thabtah, O. Gharaibeh, and R. Al-Zubaidy, "Arabic Text Mining using Rule based Classification", *Journal of Information and Knowledge Management,* Vol. 11, No. 01, 2012.

[22] A.S. Ghareb, A.R. Hamdan, and A.A. Bakar, "Integrating Noun-Based Feature Ranking and Selection Methods with Arabic Text Associative Classification", Arabian Journal for Science and Engineering, Springer, Berlin Heidelberg, 2014.
http://link.springer.com/article/10.1007/s13369-014-1304-3

[23] A.S. Ghareb, A.R. Hamdan, and A.A. Bakar, "Text Associative Classification approach for mining Arabic data set", *In: Proceeding of 4th DMO international conference on Data Mining and Optimization, IEEE*, Malaysia, September 2-4, 2012, pp. 114-120.

[24] D. Isa, V. P. Kallimani, and L.H. Lee, "Using the self organizing map for clustering of text documents", *Expert Systems with Applications*, Vol. 36, No. 5, 2009, pp. 9584-9591.

[25] D. Isa, L.H. lee, V.P. Kallimani, and R. RajKumar, "Text Documents Preprocessing with the Bahes Formula for Classification using the Support vector machine", *IEEE Transaction of Knowledge and Data Engineering*, Vol. 20, No. 0-9, 2008, pp. 1264-1272.

[26] A. Zelaia, I. Alegria, O. Arregi, and B. Sierra, "A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension", *Applied Soft Computing,* Vol. 11, 2011, pp. 4981–4990.

[27] D. Miao, Q. Duan, and H. Zhang, "Rough set based hybrid algorithm for text classification", *Expert Systems with Applications*, Vol. 36, 2009, pp. 9168–9174.

[28] W. Li, D. Miao, and W. Wang, "Two-level hierarchical combination method for text classification", *Expert Systems with Applications*, Vol. 38, 2011, pp. 2030–2039.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[29] L.I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms", Wiley, 2004.

[30] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *The Journal of Machine Learning Research,* Vol. 3, 2003, pp. 1289-1305.

[31] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in automated text categorization", *Research and Advanced Technology for Digital Libraries,* Springer, 2000, pp. 59-68.

[32] S.S. Mengle, and N. Goharian, "Ambiguity measure feature selection algorithm", *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 5, 2009, pp. 1037-1050.

[33] D. Mladenic, and M. Grobelnik, "Feature selection for unbalanced class distribution and Naive Bayes", *In: Proceeding of 16th International Conference on Machine Learning (ICML),* 1999, pp. 258-267.

[34] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information processing and management,* Vol. 24, No. 5, 1988, pp. 513-523.

[35] B. M. Zahran, and G. Kanaan, "Text Feature Selection using Particle Swarm Optimization Algorithm", *World Applied Sciences Journal (Special Issue of Computer and IT),* 7, 2009, pp. 69-74.

[36] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization", *Information Processing and Management,* Vol. 48, No. 4, 2012, pp. 741-754.

[37] A. McCallum, and K. Nigam, "A comparison of event models for Naive Bayes text classification", *AAAI-98 workshop on learning for text categorization*, 1998, pp. 41-48.

[38] F. Thabtah, "A review of associative classification mining", *The Knowledge Engineering Review,* Vol. 22, No. 01, 2007, pp. 37-65.

[39] H.K. Chantar, and D.W. Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN", *Third World Congress on Nature and Biologically Inspired Computing NaBIC*, 2011, pp. 546-551.

[40] M. Abbas, K. Smaili, and D. Berkani, "Evaluation of Topic Identification Methods on Arabic Corpora", *Journal of Digital Information Management,* Vol. 9, No. 5, 2011, pp. 185-192.