Efficient Text Classifier Using Rough Sets and Hybrid Classifier Approach: A Case Study in Elfagr Newspaper

Mohamed Omran[1], O.E. Emam[2], Laila Abd-Elatif [3], M. Thabet[4]

[1]Mathematics Department, Sadat Academy, Egypt

[2]Information Systems Department, Faculty of Computers and Information Systems, Helwan University, Egypt

[3]Information Technology Department, Faculty of Computers and Information Systems, Helwan University, Egypt

[4]Information System Department, Faculty of Computers and Information Systems, Fayoum University, Egypt

Abstract: ***Text classification is the process of classifying documents into predefined categories based on their content. It is the automated assignment of natural language texts to predefined categories. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data. Existing supervised learning algorithms for classifying text need sufficient documents to learn accurately. This paper presents an algorithm based on rough set for the automatic grouping of PDF documents, and with potential application for Web document classification.***

Keywords: rough sets, classifier, elfagr newspaper

1. Introduction

Text classification Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data also it is the process of classifying documents into predefined categories based on their content. It is the automated assignment of natural language texts to predefined categories, there are numerous text documents available in electronic form. More and more are becoming available every day. Such documents represent a massive amount of information that is easily accessible. Seeking value in this huge collection, organization requires much work to organize documents, but this can be automated through data mining as an artificial intelligence technique where the task of data mining is to automatically classify documents into predefined classes based on their content. Many algorithms have been developed to deal with automatic text classification,the most common techniques used for this purpose including naïve Bayes classifier, association rule mining, genetic algorithm, decision tree etc. Association rule mining finds interesting association or correlation among a large set of data items. The discovery of these relationships among huge amounts of transaction records can help in many decision making process

Existing supervised learning algorithms for classifying text need sufficient documents to learn accurately. This paper presents a new algorithm for text classification using artificial intelligence technique that requires fewer documents for training. Instead of using words, word relation i.e. association rules from these words is used to derive feature set from pre-classified text documents. The concept of naive Bayes classifier is then used on derived features and. A system based on the proposed algorithm has been implemented and tested.

Text classification (TC) is an important part of text mining, looked to be that of manually building automatic TC systems by means of knowledge-engineering techniques, i.e. manually defining a set of logical rules that convert expert knowledge on how to classify documents under the given set of categories. For example would be to automatically label each incoming news story with a topic like “sports”, “politics”, or “art”. a data mining classification task starts with a training set $D = (d1..... dn)$ of documents that are already labeled with a class C1,C2 (e.g. sport, politics). The task is then to determine a classification model which is able to assign the correct class to a new document d of the domain Text classification has two flavors as single label and multi-label .single label document is belongs to only one class and multi label document may be belong to more than one classes

In this paper we present a new algorithm for text classification. Instead of using words, word relation i.e. association rules is used to derive feature set from pre-classified text documents. The concept of naïve Bayes classifier is then used on derived features and finally a concept of genetic algorithm has been added for final classification. A system based on the proposed algorithm has been implemented and tested. The experimental results show that the proposed system works as a successful text classifier.

2. Text Classification Process

Main TC Stages:

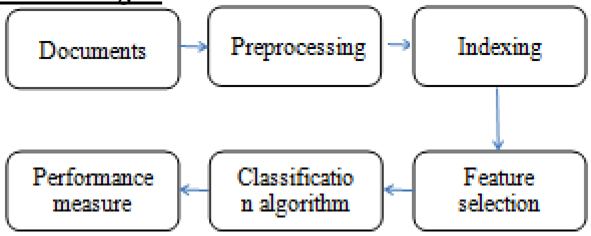


Figure 1: TC Stages

Documents Collection

This is first step of classification process in which we are collecting the different types (format) of document like html, .pdf, .doc, web content etc

Pre-Processing

The first step of pre-processing which is used to presents the text documents into clear word format. The documents prepared for next step in text classification are represented by a great amount of features. Commonly the steps taken are:

- ***Tokenization***: A document is treated as a string, and then partitioned into a list of tokens. Removing stop words: Stop words such as “the”, “a”, “and”, etc are frequently occurring, so the insignificant words need to be removed.
- ***Stemming word***: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form.

Indexing

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector The Perhaps most commonly used document representation is called vector space model (SMART) [1] . HTML tags are used to build the web document representation.

Feature Selection

After pre-processing and indexing the important step of text classification, is feature selection [2] to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The main idea of Feature Selection (FS) is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word. Because of for text classification a major problem is the high dimensionality of the feature space.

Classification

The automatic classification of documents into predefined categories has observed as an active attention, the documents can be classified by three ways, unsupervised, supervised and semi supervised methods. From last few years, the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks.

Performance Evaluations

This is Last stage of Text classification, in which the evaluations of text classifiers is typically conducted experimentally, rather than analytically. The experimental evaluation of classifiers, rather than concentrating on issues of Efficiency, usually tries to evaluate the effectiveness of a classifier, i.e. its capability of taking the right categorization decisions. An important issue of Text categorization is how to measures the performance of the classifiers. Many measures have been used, like Precision and recall [3]; fallout, error, accuracy etc.

3. Classifiers

1) Rocchio’s Algorithm

Rocchio’s learning algorithm [4] is in the classical IR tradition. It was originally designed to use relevance feedback in querying full-text databases, Rocchio’s Algorithm is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class ci , and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

2) K-Nearest Neighbors

K-NN classifier is a case-based learning [5] algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measure’s This method is try for many application [6] Because of its effectiveness, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k .The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct

3) Naïve Bayes

Naïve bias method is kind of module classifier [7] under known priori probability and class conditional probability .it is basic idea isto calculate the probability that document D is belongs to class C. There are two event model are present for naive Bias [8] as multivariate Bernoulli and multinomial model. Out of these model multinomial model is more suitable when database is large, but there are identifies two serious problem with multinomial model first it is rough parameter estimated and problem it lies in handling rare categories that contain only few training documents

4) Decision tree

When decision tree is used for text classification it consist tree internal node are label by term, branches departing from them are labeled by test on the weight, and leaf node are represent corresponding class labels .Tree can classify the document by running through the query structure from root to until it reaches a certain leaf, which represents the goal for the classification of the document. Most of training data will not fit in memory decision tree construction it becomes inefficient due to swapping of training tuples. To handle this issue [9] presents method which can handle numeric and categorical data.

5) SVM

The application of Support vector machine (SVM) method to Text Classification has been propose by [10]. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector.

6) Neural Network

A neural network classifier is a network of units, where the input units usually represent terms, the output unit(s) represents the category. For classifying a test document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit(s) takes up as a consequence determines the categorization decision. Some of the researches use the single-layer perceptron, due to its simplicity of implementing . The multi-layer perceptron which is more sophisticated, also widely implemented for classification tasks .Models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [11] for documents classification

7) LLSF

LLSF stands for Linear Least Squares Fit, a mapping approach developed by Yang [12]. The training data are represented in the form of input/output vector pairs where the input vector is a document in the conventional vector space model (consisting of words with weights), and output vector consists of categories (with binary weights) of the corresponding document. Basically this method is used for Information Retrieval [13] for giving the output of query in form of relevant document but it can easily use for text classification

8) Associative classifier

Recent studies in the data mining community proposed new methods for classification employing association rule mining. These associative classifiers have proven to be powerful and achieve high accuracy. [14]. The main idea behind this algorithm is to scan the transactional database searching for k-item sets relationships among items in a transactional database To Build an Associative Text Classifier construction phases are shown in following figure.

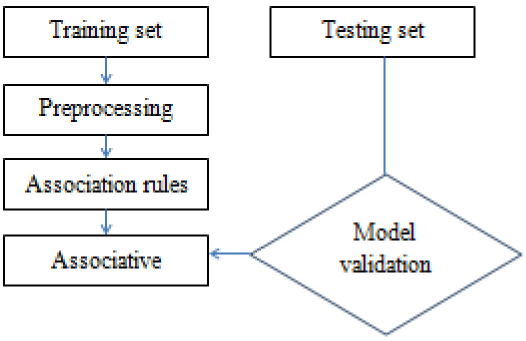


Figure 2: Associative Classifier

The first three steps belong to the training process while the last one represents the testing (or classification) phase. More details on the process are given in the subsections below Collecting the training set document after performing the pre-processing; in second phase using association algorithm on the documents would generate a very large number of association rules

4. Proposed Algorithm

The proposed method for classifying text is an implementation of a hybrid method consisting of association rule, naïve Bayes classifier, and genetic algorithm. The features of association rule are used to make association sets. On the other hand, to make a probability chart with prior probabilities, naïve Bayes classifier's probability measurements are used. And finally in the retrieval phase we have implemented the positive-negative matching calculation found in different researches of genetic algorithm [15], [16]. Here the associated word sets, which do not match with considered class is treated as negative sets and others are positive.

The following algorithm is used for class determination in testing phase.

```
 n = number of class
m = number of associated sets
1. for each class i = 1 to n do
2. setpval = 0, nval = 0, p = 0, n = 0
3. for each set s = 1 to m do
4. if the probability of the class (i) for the set (s) is maximum then increment pval else increment nval
5. if 50% of the associated set s is matched with the keywords set do step 6 else do step7
6. if maximum probability matches the class i then increment p
7. if maximum probability does not match the class i increment n
8. if (s<=m) go to step 3
9. calculate the percentage of matching in positive sets for the class i
10. calculate the percentage of not matching in negative sets for the class i
11. calculate the total probability as the summation of the results obtained from step 9 and 10 and also the prior probability of the class i in set s
12. if (i<=n) go to step 1
13. set the class having the maximum probability value as the result.
```

Dataset Features

Dataset was taken at the period of May 2017 directly from AL-Fajr databases. A total number of dataset records are 50,000.

Dataset Attributes

Dataset was collected from AL-Fajr databases and has the following attributes:

- **CatID:** refer to category id with a number data type.
- **SubCatID:** refer to sub category id with a number data type.
- **PhotoState:** refer to state of the photo with a number data type.

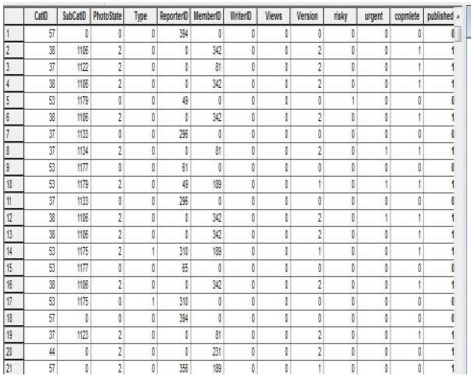
- **ReporterID:** refer to id of the reporter with a number data type.
- **MemberID:** refer to id of the member with a number data type.
- **Views:** refer to count of the views with a number data type.
- **Risky:** refer to risky degree with a number data type.
- **Urgent:** refer to urgent degree with a number data type.
- **Complete:** refer to complete degree with a number data type.
- **Version:** refer to version number with a number data type.
- **Published:** refer to whether the new is published or no.

News Analysis and Building Decision Rules Based on Rough Large Scale Integer Linear Programming problem

4.1 Information Table Construction

The normalized dataset represented the IS of the news analysis, which includes the following:

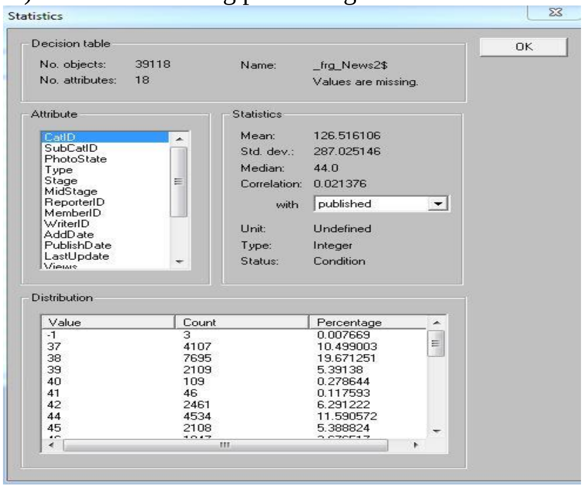
- Condition attributes are {CatID, SubCatID, PhotoState, ReporterID, MemberID, Views, Risky, Urgent, Complete, Version}.
- Decision attribute is {Published} which has two values. If valued "Yes," then the new has been published; if valued "No," then the new has not been published.



	CatID	SubCatID	PhotoState	Type	ReporterID	MemberID	WriterID	Views	Version	risky	urgent	copmlete	published
1	57	0	0	0	394	0	0	0	0	0	0	0	0
2	38	1106	2	0	0	342	0	0	2	0	0	1	1
3	37	1122	2	0	0	81	0	0	2	0	0	1	1
4	38	1106	2	0	0	342	0	0	2	0	0	1	1
5	53	1179	0	0	49	0	0	0	0	1	0	0	0
6	38	1106	2	0	0	342	0	0	2	0	0	1	1
7	37	1133	0	0	296	0	0	0	0	0	0	0	0
8	37	1134	2	0	0	81	0	0	2	0	1	1	1
9	53	1177	0	0	61	0	0	0	0	0	0	0	0
10	53	1179	2	0	49	189	0	0	1	0	1	1	1
11	37	1133	0	0	296	0	0	0	0	0	0	0	0
12	38	1106	2	0	0	342	0	0	2	0	1	1	1
13	38	1106	2	0	0	342	0	0	2	0	0	1	1
14	53	1175	2	1	310	189	0	0	1	0	0	1	1
15	53	1177	0	0	65	0	0	0	0	0	0	0	0
16	38	1106	2	0	0	342	0	0	2	0	0	1	1
17	53	1175	0	1	310	0	0	0	0	0	0	0	0
18	57	0	0	0	394	0	0	0	0	0	0	0	0
19	37	1123	2	0	0	81	0	0	2	0	0	1	1
20	44	0	2	0	0	231	0	0	2	0	0	0	1
21	57	0	2	0	356	189	0	0	1	0	0	0	1

Figure 3: News Analysis Representation

Figure 3 represents the News Analysis in the form of condition attributes and decision attribute (Information Table) for further coming processing.



Statistics

Decision table

No. objects: 39118
No. attributes: 18
Name: _frg_News2$
Values are missing.

Attribute

CatID
SubCatID
PhotoState
Type
Stage
MidStage
ReporterID
MemberID
WriterID
AddDate
PublishDate
LastUpdate
Views

Statistics

Mean: 126.516106
Std. dev.: 287.025146
Median: 44.0
Correlation: 0.021376
with published
Unit: Undefined
Type: Integer
Status: Condition

Distribution

Value	Count	Percentage
-1	3	0.007669
37	4107	10.499003
38	7695	19.671251
39	2109	5.39138
40	109	0.278644
41	46	0.117593
42	2461	6.291222
44	4534	11.590572
45	2108	5.388824

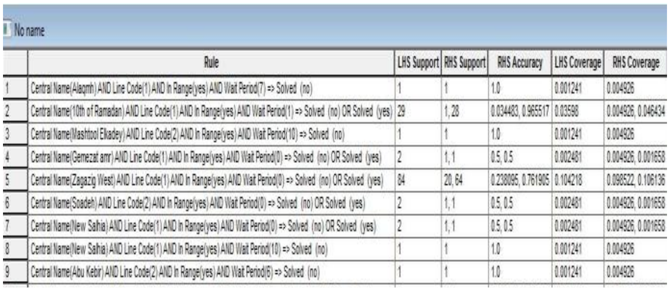
Figure 4: Domain Values of the News Analysis Condition Attributes

In Figure 4, the Rosetta shows the domain values of the condition attributes and the decision attribute of the News Analysis, number of occurrences and the coverage percentage of each value.

4.2 Building Decision Rules for the News Analysis

In Figure 8, Rosetta will be used to build decision rules of the News Analysis into the form (LHS→ RHS) and consider the following for each rule:

- **LHS Support:** No. of news that satisfy condition attributes in the LHS of the rule.
- **RHS Support:** No. of news that satisfy decision attribute in the RHS of the rule.
- **RHS Accuracy:** No. of news that satisfy condition attributes in the LHS of the rule by No. of news that satisfy decision attribute in the RHS of the rule.
- **LHS Coverage:** No. of news that satisfy condition attributes in the LHS of the rule by total No. of news.
- **RHS Coverage:** No. of news that satisfy decision attribute in the RHS of the rule by total No. of news.



No name

	Rule	LHS Support	RHS Support	RHS Accuracy	LHS Coverage	RHS Coverage
1	Central Name(Alaqmh) AND Line Code(1) AND In Range(yes) AND Wait Period(7) => Solved (no)	1	1	1.0	0.001241	0.004926
2	Central Name(10th of Ramadan) AND Line Code(1) AND In Range(yes) AND Wait Period(1) => Solved (no) OR Solved (yes)	29	1, 28	0.034483, 0.965517	0.03598	0.004926, 0.046434
3	Central Name(Mashtool Elkadey) AND Line Code(2) AND In Range(yes) AND Wait Period(10) => Solved (no)	1	1	1.0	0.001241	0.004926
4	Central Name(Gemezat amr) AND Line Code(1) AND In Range(yes) AND Wait Period(0) => Solved (no) OR Solved (yes)	2	1, 1	0.5, 0.5	0.002481	0.004926, 0.001658
5	Central Name(Zagazig West) AND Line Code(1) AND In Range(yes) AND Wait Period(0) => Solved (no) OR Solved (yes)	84	20, 64	0.238095, 0.761905	0.104218	0.098522, 0.106136
6	Central Name(Soadeh) AND Line Code(2) AND In Range(yes) AND Wait Period(0) => Solved (no) OR Solved (yes)	2	1, 1	0.5, 0.5	0.002481	0.004926, 0.001658
7	Central Name(New Salhia) AND Line Code(1) AND In Range(yes) AND Wait Period(0) => Solved (no) OR Solved (yes)	2	1, 1	0.5, 0.5	0.002481	0.004926, 0.001658
8	Central Name(New Salhia) AND Line Code(1) AND In Range(yes) AND Wait Period(10) => Solved (no)	1	1	1.0	0.001241	0.004926
9	Central Name(Abu Kebir) AND Line Code(2) AND In Range(yes) AND Wait Period(6) => Solved (no)	1	1	1.0	0.001241	0.004926

Figure 5: News Analysis Decision Rules Using Rosetta

4.3 Operational Research Algorithms in Handling News Analysis

Emam [17] et al. focused on the solution of fully rough three level large scale integer linear programming problem, in which all decision parameters and decision variables in the objective functions and the constraints are rough intervals, and have block angular structure of the constraints. This paper based on block angular structure where news analysis is distributed among several multiple departments, each department has its own constraints and communicate with other departments through common constraints which represent the linked point between all departments.

News Analysis can be modeled as "Rough Large Scale Integer Linear Programming" problem based on Operational Research (OR) techniques, where:

- It is distributed among several multiple sub problems (departments).
- Decision rules of each department represent the constraints of this department.
- Constraints of each department are independent of others.
- No. of news which can be published are integer values.
- There are a set of news which certainly can be published (Lower Approximation) and a set of news which possibly can be published (Upper Approximation).

4.4 News Analysis Modeling In OR Formulation

In the first, we transform the decision rules of each department into equations.

In the second, we formulate the common constraints which applied to all departments.

Finally, we create the objective function which represent the status of news will be checked.

And, we will get the following formula:

Max Objective Function,
Subject to
Common Constraint,

Independent constraints of each department which contain rough values in the right hand side of the equation.

To solve the problem, we will decompose the problem into two parts.

The first problem formulated when putting the value of right hand side of equations by "Yes".

The second problem formulated when putting the value of right hand side of equations by "No".

4.5 News Analysis in the View of Positive Region, Boundary Region, and the Outside Region

The “Published” attribute valued “Yes,” and the following three regions are considered:

- **Lower Approximation (Positive Region):** Set of news which certainly can be published (47702).
- **Boundary Region:** Set of news which possibly can be published.
- **Outside Region:** Set of news which certainly can’t be published.
- **The accuracy of Approximation** = 47702/49341=0.969, which means 96.6% of news with “Published” attribute valued “Yes” are certainly published and 3.4% of its are possibly published.

The “Published” attribute valued “No,” and the following three regions are considered:

- **Lower Approximation (Positive Region):** Set of news which certainly can’t be published (659).
- **Boundary Region:** Set of news which possibly can be published.
- **Outside Region:** Set of news which certainly can be published.
- **The accuracy of Approximation** = 659/2289=0.286, which means 28.6% of news with “Published” attribute valued “No” are certainly can’t be published and 71.4% of its are possibly published.

References

[1] [KjerstiAas and Line Eikvil “Text Categorization: A Survey” Report No. 941. ISBN 82-539-0425-8. , June, 1999

[2] A. Dasgupta, P. Drineas, B. Harb, “Feature Selection Methods for Text Classification”, KDD’07, ACM, 2007

[3] Yiming Yang “An Evolution of statistical Approaches to Text Categorization” Information Retrieval 1, 69-90 1999

[4] Hein Ragas Cornelis H.A. Koster, “Four text classification algorithms compared on a Dutch corpus” SIGIR 1998: 369-370 1998

[5] GongdeGuo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, “KNN Model-Based Approach in Classification”, Proc. ODBASE pp- 986 – 996, 2003

[6] EijiAramaki and Kengo Miyo, “Patient status classification by using rule based sentence extraction and bm25-knn based classifier”, Proc. of i2b2 AMIA workshop, 2006

[7] SHI Yong-feng, ZHAO, “Comparison of text categorization algorithm”, Wuhan university Journal of natural sciences. 2004

[8] [Vidhya. K.A G.Aghila, “A Survey of Naïve Bayes Machine Learning approach in Text Document Classification”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010

[9] Mnish Mehta, Rakeshagrwal” SLIQ: A Fast Scalable Classifier for Data Mining” 1996

[10] Joachims, T. “Text categorization with support vector machines: learning with many relevant features”. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, DE), pp. 137–142 1998

[11] Cheng HuaLi , Soon Choel Park “An efficient document classification model using an improved back propagation neural network and singular value decomposition”, Expert Systems with Applications, 3208–3215, 2009

[12] Yiming Yang And Christopher G. Chute Mayo Cllnic “An Example-Based Mapping Method For Text Categorization And Retrieval” ACM Transactions On Information Systems, Vol. 12, No 3, Pages 252-277, July 1994

[13] Yiming Yang Christopher G. Chute “A Linear Least Squares Fit Mapping Method For Information Retrieval From Natural Language Texts” Acres De Coling-92 Nantes, 23-28 AOUT 1992

[14] O. Zaiane, and M. Antonie, “Text Document Categorization by Term Associaton”, Proceedings of ICDM 2002, IEEE, , pp.19-26 2002

[15] Anwar M. Hossain, Mamunur M. Rashid, ChowdhuryMofizurRahman, “A New Genetic Algorithm Based Text Classifier,” In Proceedings of International Conference on Computer and Information Technology, NSU, 2001, pp. 135-139.

[16] EshitaSharmin, Ayesha Akhter, ChowdhuryMofizurRahman, “Genetic Algorithm for Text Categorization,” In Proceedings of International Conference on Computer and Information Technology, BUET, December, 1998, pp. 80-85.