

Learning to Filter Text in Forum Malay Message using Naïve Bayesian Technique

By

Norhadila Binti Ab. Halim

Dissertation submitted in partial fulfillment of
the requirements for the
Bachelor of Technology (Hons)
(Information Communication Technology)

JANUARY 2006

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

t
QA
279.5
.N577
2006

~~1) Bayesian statistical decision theory
2) ST/IS - theory~~

CERTIFICATION OF APPROVAL

Learning to Filter Text in Forum Malay Message
Using Naïve Bayesian Technique

By

Norhadila Ab Halim

Dissertation Submitted to the Information Technology Programme
Universiti Teknologi PETRONAS
In partial fulfillment of the requirement for the
Bachelor of Technology (Hons)
(Information Communication Technology (ICT))

Approved By,



(Ms. Norshuhani Bt Zamin)

UNIVERSITI TEKNOLOGI PETRONAS
TRONOH, PERAK

Jan 2006

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.



NORHADILA BINTI AB. HALIM

ABSTRACT

Applying the basic filtering technique in forum application has been discussed in [1]. The paper explains about the use of the basic naïve Bayesian algorithm to classify forum messages whether clean or bad where clean message has no bad words, while bad message contains at least one bad word. In this Final Year Project paper, the application of the algorithm in the filtering forum messages will be discussed in the attempt to apply learning to filter forum messages.

ACKNOWLEDGEMENT

In the name of Allah s.w.t, the Most Gracious, the Most Merciful.

First and foremost, the author would like to express her greatest gratitude to Allah the Almighty for His guidance and help during the course of life and moment of truth. Heartiest thanks firstly go to the most dedicated person, Norshuhani Binti Zamin for her kindness and guidance in completing the project. Thank you again for your endless encouragement and patience.

The author also would like to express her deepest appreciation and sincere gratitude to her beloved friends and colleagues namely Azlina Binti Daud, and Sarah Haryati Zulkifli for the willingness and greatest cooperation to share their experiences and knowledge through thin and thick. Not to forget, the author would like to acknowledge this wonderful person Nurizza for assisting her in understanding the Bayesian concept and implementation. Greatest thanks again to all for the continual support and assistance. Author's big gratitude also goes to Universiti Teknologi Petronas (UTP) for providing convenience place and superb facilities for learning.

Last but not least, the author would like to express a special thank to her family for their understanding, priceless support and constant love throughout the accomplishment of the project. Lastly, thank you again to all people who involved directly or indirectly starting from the first day until the end of this project completion. Your endless contributions are highly appreciated.

TABLE OF CONTENTS

CERTIFICATION		i-ii
ABSTRACT		iii
ACKNOWLEDGEMENT		iv
CHAPTER 1:	INTRODUCTION	1 - 3
	1.1 Background of Study	1
	1.2 Problem Statement	2
	1.3 Objective	3
	1.4 Scope of Study	3
CHAPTER 2:	LITERATURE REVIEW	4 - 6
	2.1 History and Previous Works.	4 - 6
CHAPTER 3:	METHODOLOGY	7-12
	3.1 Project Development Phases	7
	3.1.1 Research and Review	8
	3.1.2 Conceptualization	8
	3.1.3 Knowledge Acquisition and Analysis	8
	3.1.4 Design and Implementation.	8
	3.1.5 Testing	9
	3.1.6 Documentation & Management	9
	3.2 System Architecture	9-10
	3.3 Main Method and Technique	10
	3.4 Tools Required	13
	3.5.1 Hardware	13
	3.5.2 Software	13

CHAPTER 4:	RESULT AND DISCUSSION	14-15
	4.1 Screen shots	14
	4.2 System Limitations.	15
	4.3 Suggestion for future works.	15
CHAPTER 5:	CONCLUSION AND RECOMMENDATION.	16
REFERENCES		17-20

LIST OF FIGURES

Figure 3.1	System's Development Methodology	7
Figure 3.2	Stand alone	9
Figure 3.2.1	System Flow	10
Figure 4.1	Screen shot 1	13

CHAPTER 1

INTRODUCTION

1.1 Background of study

The explosive growth of the Internet and other sources of networked information have made automatic mediation of access to networked information sources an increasingly important problem. Much of this information is expressed as electronic text, and it is becoming practical to automatically convert some printed documents and recorded speech into electronic text as well. Thus, automated systems capable of detecting useful documents are finding widespread application. One important type of automated text detection system is called a text filtering system. In this project, an automated text detection system or a text filtering system is used to filter text in forum message. Forum can be defined as an online discussion group, where participants with common interests can exchange open messages. Forum messages can be varied according to the forum focus area whereby they can be something related to computer, business, or other areas. Obviously, forum is a place where people with the same area of interest meet online, discuss and change ideas among them.

There are many sources that discussed about text filtering especially in e-mail application for example the anti-spam filtering [4]. The basic idea of this project is to develop an application to filter Malay language text in forum message by classifying clean and bad messages using Naïve Bayesian algorithm. Naïve Bayesian algorithm has been a popular technique in text filtering as it offers potential Bayesian classifier that can produced accurate and reliable results in the end [11] [12].

1.2 Problem Statement

People join forum to share ideas on certain topic or key areas which they are interested in. It can be a positive discussion as well as negative ones. Sometimes the forum users do not realize whether the messages are the bad or clean messages. The administrator will have heavy workloads if he wants to classify the bad and the clean messages on a certain date and time. The problem comes when there are forum users who illegally misuse the words or vocabulary which can lead to confusion and misunderstanding. In fact, forum that allows it users to use Malay language always face the same problems when forum users tend to use words which contain harsh meaning and sometimes even critical to be displayed for public view. In addition, the number of this application in this area is not widely implemented yet whereby there's only few text filtering system in forum message exist. On the other hand, the existence of this system can allow the administrator to analyze the trend or message pattern in forum. Before this, they archived all forum messages without any intention to know how many of them are bad and clean messages.

1.3 Objectives

1. To learn in filtering text in forum messages using Naïve Bayesian algorithm with special focus on Malay language.
2. To provide a text filtering system in forum message which classify the messages into two categories; bad and clean messages. Clean messages contain no bad words while bad messages have one or more bad words.
3. To provide an effective and accurate filtering application for forum messages.
4. To apply Naïve Bayesian algorithm in text filtering system that can be used in any kind of digital text document.

1.4 Scope of study

At the moment, this project is aimed to filter text in forum messages with special focus on Malay language messages. Its last result should be in two categories which are bad messages and clean messages. Clean messages have no bad words and on the other hand, bad messages contain at least one bad word. Examples of bad words would be *sial*, *bodoh*, *gila*, *bangang* that can be found in the forum messages. Hopefully in the future, this project can be enhanced by focusing on both, English and Malay messages.

CHAPTER 2

LITERATURE REVIEW

2.1 History and Previous Works

So-called 'naïve' Bayesian classification is the optimal method of supervised learning if the values of the attributes of an example are independent given the class of example. Although this assumption is almost always violated in practice, recent work has shown that naïve Bayesian learning is remarkably effective in practice and difficult to improve upon systematically [12]. Naïve Bayesian learning gives better test set accuracy than any other known method, including back propagation and C4.5 decision trees [13]. According to Lang [2], Bayesian Learning just reduces the probability of an inconsistent hypothesis. This gives the Bayesian Learning a bigger flexibility. The Bayesian Learning Algorithms combine training data with a priori knowledge to get a posterior probability of a hypothesis. So it is possible to figure out the most probable hypothesis according to the training data.

Referring to [9], a naive Bayes classifier (also known as Idiot's Bayes) is a simple probabilistic classifier. Naive Bayes classifiers are based on probability models that incorporate strong independence assumptions which often have no bearing in reality, hence are (deliberately) naive. A more descriptive term for the underlying probability model would be *independent feature model*. Furthermore the probability model can be derived using Bayes' theorem. Meanwhile in order to improve Naïve Bayesian classifier, Shen et. al [3] suggested Naive Bayesian is trained as a "generative" model that fits the distribution of the data instances given the class label; the method is to add a small number of parameters that are trained like a "discriminative model," and fits the

distribution of the class label given the instance. Focusing on the problem of text classification, by adding this discriminative component to naive Bayesian significantly lowers the test error and leads to much improved accuracy/coverage curves.

Filtering system has become a popular application since many techniques are brought by researchers in order to achieve the most accurate ones. Michelakis et. al [6] claimed in his paper that Filtron, learning-based anti-spam filters, mostly based on Naive Bayes, are becoming operational. This paper presents Filtron, a prototype anti-spam filter implementation that incorporates the key findings of our previous work. Filtron emerged as the result of our thorough investigation of learning approaches to anti-spam filtering. Sahami et. al [7] found that a rule-based approach is of limited utility in junk mail filtering. This is due to the fact that such logical rule sets usually make rigid binary decisions as to whether to classify a given message as junk. These rules generally provide no sense of a continuous degree of confidence with which the classification is made. Such a confidence score is crucial if we are to consider the notion of differential loss in misclassifying Email. Since the cost of misclassifying a legitimate message as junk is usually much higher than the cost of classifying a piece of junk mail as legitimate, a notion of utility modeling is imperative. To this end they require firstly a classification scheme that provides a probability for its classification decision and second some quantification of the difference in cost between the two types of errors in this task. Given these, it becomes possible to classify junk Email within a Decision Theoretic framework. There has recently been a good deal of work in automatically generating probabilistic text classification models such as the Naive Bayesian classifier.

Androutsopoulos et. al [4] in his research paper evaluated Naïve Bayesian for anti-spam filtering system. The research group investigated on the effect of attribute-set-size, training-corpus size, lemmatization, and stop lists on the filter's performance and some issues that had not been previously explored. As a conclusion from the research, additional safety nets are needed for the Naïve Bayesian anti-spam filter to be viable in

practice. Moreover, Gee [15] suggested the used of latent semantic indexing in filtering spam email. His paper analyzed the effectiveness of another machine learning approach, latent semantic indexing to the problem of filtering spam and legitimate email. Latent semantic indexing is a statistical technique that derives correlations between terms and documents in a corpus and reflects indirect inferential. Latent semantic indexing (LSI) is a statistical technique that derives a statistical correlation between all terms and documents in a corpus, in an attempt to overcome the problems inherent in lexical matching.

Sainin [1] in a paper called “Learning to Filter Text in Forum Message” (2005) explained about the use of naïve Bayesian algorithm to classify forum messages whether clean or bad where clean message has no bad words, and bad message contains one or more bad words. It also discussed on the modification of the algorithm including pre-processing and classification in the attempt to apply learning to filter form messages. According to the paper, the development of idea was based on Artificial Intelligence Special Interest Group (AISIG) e-Community Portal that is currently using an anonymous message submission in eJava Forum. It is to encourage student and visitors to ask Java related questions and at the same time response to the forum without registering their username and password. This filtering application can support two languages namely Malay and English words. There are two basic components in this system which are pre-processing and classifier model that was developed to experiment the learning and filter the text in forum messages. The preprocessing task is to clean the text messages from the forum database and can be further used to build the model for naïve Bayes learning. Preprocessing tasks include removing repeated string and collecting all words and punctuations. The new preprocessing algorithm was also applying a method to remove stop words and common words that normally used in English and Malay language for example “and”, “the”, “yang” and “itu”. The sorted and cleaned text from the preprocessing task will be passed through LEARN_BAYES_TEXT function to learn and calculate all probabilities. Probabilities that will be calculated are probability for each target value v given the number of document with target is v and probability for each word in the vocabulary.

CHAPTER 3

METHODOLOGY

3.1 Project Development Phases

The development of this system is based on methodology that is adopted from several existing methodologies for different applications as this filtering system will be an integration of these technologies.

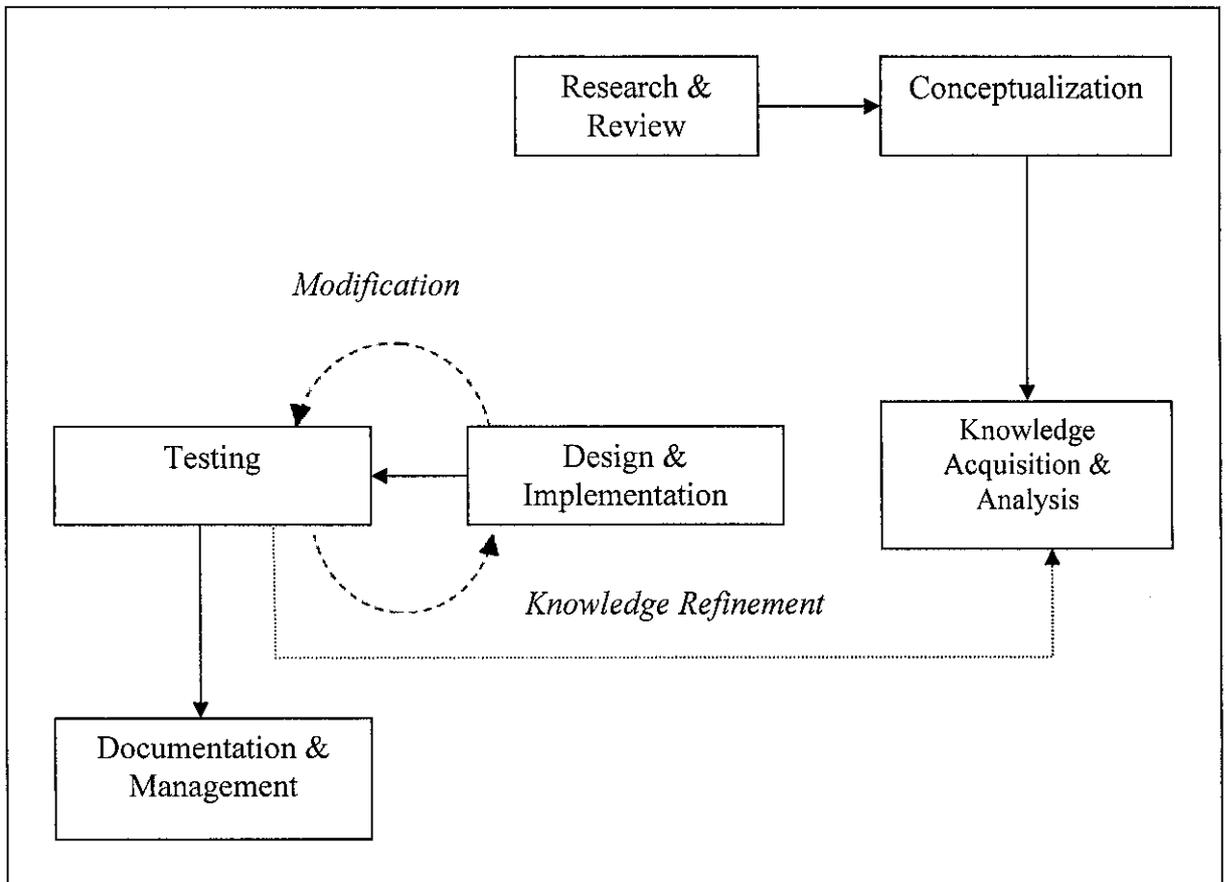


Figure 3.1: System's Development Methodology

3.1.1 Research and Review

Research has been done regarding the use of text filtering system in forum message using Naïve Bayesian algorithm. Applications available on the internet consist of many different subjects such as anti-spam filtering and content-based filtering and so on. The research would also go into reviewing the literature of underlying concepts behind the development of several application of filtering system. Results of researches and reviews conducted help in giving an idea and insight on how to implement a filtering system.

3.1.2 Conceptualization

The idea of building this filtering system and the problem domain is determined for this system. Once the problem domain is determined, it was used and set to be the title of the system development. The technologies that are needed, which includes the hardware, software, and filtering method are then selected for this system.

3.1.3 Knowledge Acquisition and Analysis

The process of acquiring knowledge in the development of a filtering system is very important. That knowledge acquired has to then be analyzed in order to find the connections and meaning between them. For this system, the knowledge was acquired from the internet, from journals and books regarding the problem domain.

3.1.4 Design and Implementation

After analyzing the information gathered, the structure and design of the system is to be implemented. In researching the appropriate method to use, Naïve Bayesian learning was chosen after comparison with several other techniques. Criteria such as accurateness and greater flexibility of algorithm were taken into account. The user

interface is designed using Visual Basic.Net. The product in this phase will be a prototype system which will be refined in later phases.

3.1.5 Testing

The prototype system is then tested to ensure that it functions accordingly. The results are compared and this phase will be repeated due to any changing and improvement. The main objective of this filtering system is to provide effective classifying text for forum message using Naïve Bayesian method. Note that all errors are also identified in this phase and corrected.

3.1.6 Documentation & Management

At the end of the development of any system, it is wise to document aspects of the development process which may come become handy later on in the future as reference for future development. Maintenance on the prototype system is to be done in order to make sure the system is reliable and updated.

3.2 System Architecture

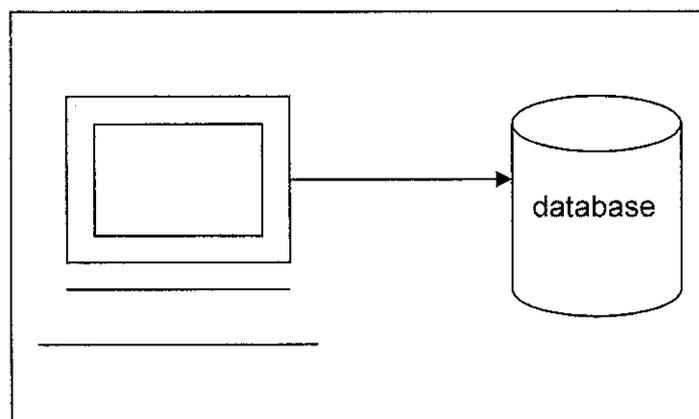


Figure 3.2: Stand alone

The database for this system is Microsoft Access 2003 and it is a stand alone window based application. Data (forum messages) are stored in the database and link to the system through ADO.Net in Visual Basic.Net

3.2.1 System Flow

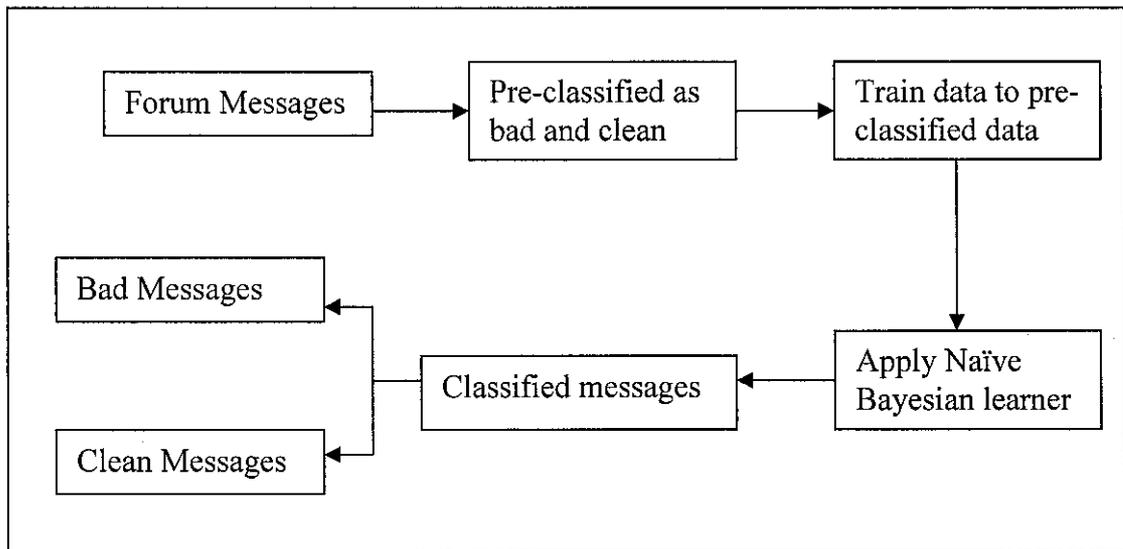


Figure 3.2.1: System flow

3.3 Main Method and Technique

This project applies Naïve Bayesian algorithm as its main technique for filtering text in forum messages.

3.3.1 Bayesian Learning

Other learning algorithms eliminate those hypotheses, which are not consistent to a training example. Whereas the Bayesian Learning just reduces the probability of an inconsistent hypothesis. This gives the Bayesian Learning a bigger flexibility. The Bayesian Learning Algorithms combine training data with a priori knowledge to get the posterior probability of a hypothesis. So it is possible to figure out the most probable

hypothesis according to the training data. The basis for all Bayesian Learning Algorithms is the Bayes Rule.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D
- P(D|h) = probability of D given h

3.3.2 Naïve Bayesian Algorithm

The learning task in naïve Bayesian classifier includes building probability estimations from set of instance x described by conjunction of attribute values and some finite target function $f(x)$. Furthermore, naïve Bayesian learner classification task is to predict the value for the new instance described by tuple of attributes $\langle a_1, a_2, \dots, a_n \rangle$. Given with a set of target value V , Bayesian approach is to classify the new instance with the most probable target value, *VMAP*. The value for *VMAP* can be calculated using Equation 1.

X be a set of instances $x_i = (a_1, a_2, \dots, a_n)$

V be a set of classifications v_j

Naive Bayesian assumption:

$$\begin{aligned} v &= \max_{v_j \in V} P(v_j | a_1, a_2, \dots, a_n) \\ &= \max_{v_j \in V} \frac{P(a_1, a_2, \dots, a_n | v_j)P(v_j)}{P(a_1, a_2, \dots, a_n)} \\ &= \max_{v_j \in V} P(a_1, a_2, \dots, a_n | v_j)P(v_j) \\ P(a_1, a_2, \dots, a_n | v_j) &= \prod_i P(a_i | v_j) \end{aligned}$$

This leads to the following algorithm:

Naive_Bayes_Learn (examples)

for each target value v_j

estimate $P(v_j)$

for each attribute value a_i of each attribute a

estimate $P(a_i | v_j)$

Classify_New_Instance (x)

$$v_{NB} = \arg \max_{v_j \in V} \prod_i P(a_i | v_j)$$

3.3.3 Implementation of Naïve Bayesian

Example of input

No	Forum Messages	Result
1	Setan la ko.	Bad
2	Ko neh siot la.	Bad
3	Engkau janganla buat macam ni.	Clean
4	Mari kita berkelah esok.	Clean
5	Mampos la aku cani.	Bad

- “Setan la ko”. For this kind of forum message it will be classified by naïve Bayesian learner as bad as it contains one bad word which is *setanl*.
- “Mari kita berkelah esok”. For this case, it is a clean message as it contains no bad word.

Probabilities and frequencies

Forum message	Result			
	Bad	Clean	Bad	Clean
Msg1	1/3	0/2	3/5	2/5
Msg2	1/3	0/2		
Msg3	0/3	1/2		
Msg4	0/3	1/2		
Msg5	1/3	0/2		

Assume : Classify “Setan la ko”

$$P(\text{Bad}) = 1/3 \cdot 3/5 = 0.067$$

$$P(\text{Clean}) = 0/2 \cdot 2/5 = 0$$

Therefore, maximum probability = **Bad**

3.4 Tools Required

3.4.1 Hardware

- Desktop PC
- Operating System: Microsoft Window XP Professional Service Pack 1
- Intel Pentium 4 2.67 GHz
- 512MB RAM
- 40 GB Hard Disk
- 15” inch VGA monitor

3.3.2 Software

- Visual Basic.Net
- Microsoft Access 2003

CHAPTER 4

RESULT AND DISCUSSION

4.1 Screen shot

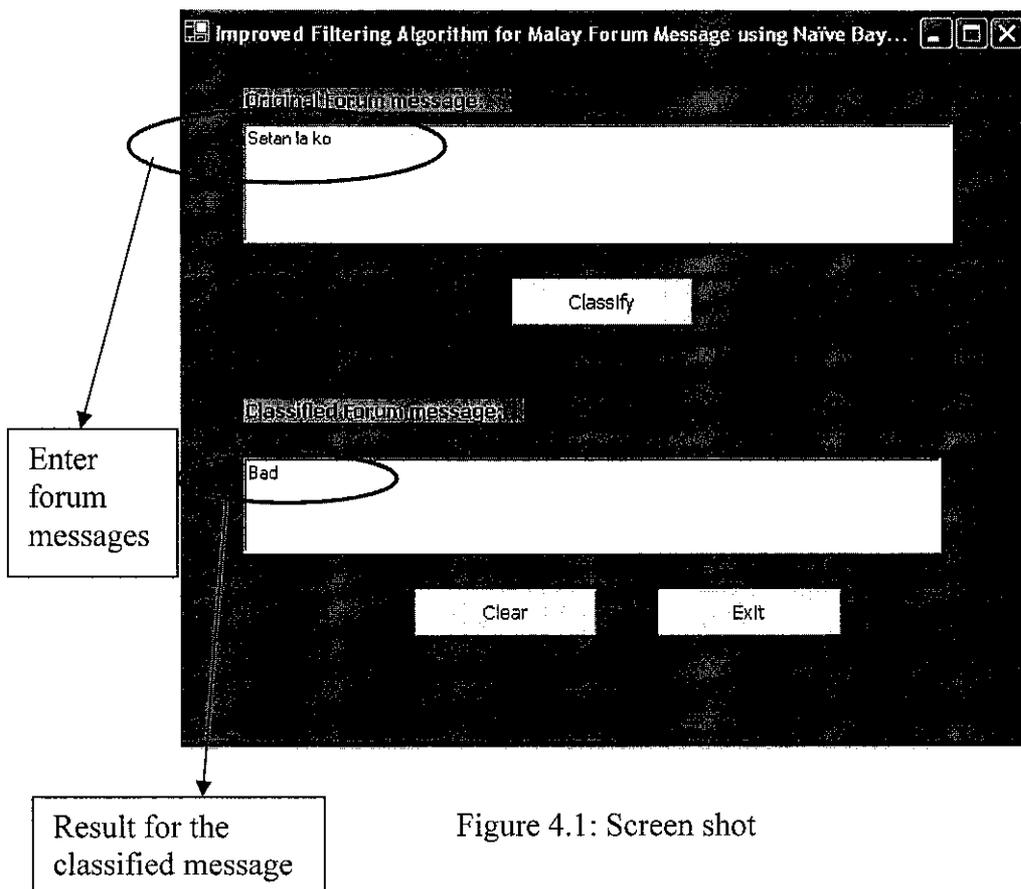


Figure 4.1: Screen shot

- **Classify** button will classify the entered message and apply naïve Bayesian at the same time in order to filter it whether it is clean or bad message.
- **Clear** button will clear the entry in both text boxes.
- **Exit** button will close the application.

4.2 System Limitations

- The number of training data is not enough to cover all words in forum messages
- Small database with small data involved decrease the accuracy
- Time constraint disallow not much improvement to be made

4.3 Suggestion for Future Works

- Modification to system approach and also Bayesian algorithm
- Better with high performance machine and improved filtering algorithm
- To expand the database with more forum messages and training data
- Cater for English and Malay forum messages
- Enhance the system for use of other digital text document

CHAPTER 5

CONCLUSION AND RECOMMENDATION

The new modification on the preprocessing algorithm contributes relatively small drop in the accuracy, however if the number of training examples is big, the search space for priors probability is expected to increase the accuracy. Again, Naïve Bayes algorithm will perform better if number of training examples is enough to cover words which are normally used in forums. There is a need to find another method to estimate the priors when the number of training examples is small and the difference between clean and bad message size is big. The experiment to apply naïve Bayes learner in forum message is interesting and the performance of the classifier including classification and processing can be improved with high performance machine and better algorithm for text filtering. Hopefully by improving the classifier it will produce more accurate result and can be used in any kind of digital text document such as blogs and others. Modern applications which applies Bayesian classifier includes the search engine Google, and the information retrieval company Autonomy Systems. They employ Bayesian principles to provide probable results to searches. Microsoft is reported as using Bayesian "probabilistic" mathematics in its future Notification Platform to filter unwanted messages. It is a very useful and beneficial algorithm that can be used in filtering any digital text document in the future.

REFERENCES

- [1] M.S Sainin , 2005 , “ Learning To Filter Text in Forum Message “, MMUSIC 2005, MMU Int. Symposium of ICT 2005, pp 13-16.
- [2] M. Lang, “ Implementation of Naive Bayesian Classifiers in Java “, Kaiserslautern University of Applied Sciences Department Zweibrücken (Germany).
- [3] Y. Shen, J. Jiang , 2003, “ Improving the Performance of Naïve Bayes for Text Classification” , CS224N Spring 2003.
- [4] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras and C. D. Spyropoulos, 2000, “ An Evaluating of Naïve Bayesian Anti-Spam Filtering”. Proceedings of the workshop on Machine Learning in the New Information Age,11th European Conference on Machine Learning, Barcelona, Spain, 2000, pp 9-17.
- [5] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, C. D. Spyropoulos, 2000, “An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages”, In Proceedings of the 23 rd Annual International ACM SIGR Conference on Research and Development in Information Retrieval, pp. 160-167.
- [6] E. Michelakis, I. Androutsopoulos, G. Paliouras, G. Sakkis and P. Stamatopoulos, 2004, “ Filtron: A Learning-Based Anti-Spam Filter”, Proceedings of the 1st Conference on Email and Anti-Spam (CEAS 2004), Mountain View, CA, USA.
- [7] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, 1998, “A Bayesian Approach to Filtering Junk E-Mail”, In AAAI Workshop, pp. 55-62.
- [8] “Bayesian Learning”, dspc11.cs.ccu.edu.tw/ml93/lecture5-bayesian-learning.pdf , 12 April 2006.

[9] “Naïve Bayes Classifier”, http://en.wikipedia.org/wiki/Naive_Bayes_classifier, 14 April 2006

[10] D.D. Lewis, 1995, “Evaluating and Optimizing Autonomous Text Classification Systems”, SIGIR’95 Sattlc WA USA 1995 ACN4 0-89791 -714-6/95 /07. S3.50

[11] P. Domingos and M. Pazani, 1996, “Beyond Independence: Conditions for the optimality of the simple Bayesian classifier”, Proceedings of the Thirteenth International Conference on Machine Learning, pages 105-112. Morgan Kaufmann Publishers, Inc., 1996.

[12] C. Elkan, 1997, “Naïve Bayesian Learning”, Department of Computer Science and Engineering, University of California, San Diego.

[13] J. S. Breese, D. Heckerman, C. Kadie, 1998, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, Technical Report MSTR-TR-98-12 Microsoft Research, Microsoft Corporation.

[14] A. S. Pannu, K. Sycara, “A Learning Personal Agent for Text Filtering and Notification”, The Robotics Institute, Schools of Computer Science, Carnegie Mellon University.

[15] K. R. Gee, 2003, “Using Latent Semantic Indexing to Filter Spam”, 2003 ACM 1-581 13-624-2/03/03.

[16] Z. Chuan, L. Xianliang, H. Mengshu, Z. Xu, “A LVQ-based neural network anti-spam email approach”, College of Computer Science and Engineering of UEST of China, Chengdu, China 610054.

[17] L. Zhang, J. Zhu, T. Yao, 2004, “An Evaluation of Statistical Spam Filtering Techniques”, ACM Transactions on Asian Language Information Processing, Vol. 3, No. 4, December 2004, Pages 243–269.

BIBLIOGRAPHY

1. "Naïve Bayes Classifier", <http://www.statsoft.com/textbook/stnaiveb.html>, 15 April 2004.
2. C. O. Brien, C. Vogel, 2002, "Spam Filters: Bayes vs. Chi-squared; Letters vs. Words", www.brightmail.com, last verified September 2002.