# Text Classification Using Naïve Bayesian Classifier with Bigram

Thandar Tin
*University of Computer Studies, Yangon*
*thandartin22@gmail.com*

## Abstract

*Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Data classification is a two step process. This system is to study the Naïve Bayesian Classifier and to classify the class labels of data sets. In this system, classifier is built on the training data sets and tests the unknown datasets. And then, calculate the accuracy of classifier by using F1-Measure (F1-score). The Naïve Bayesian (NB) classifiers have been one of the most popular techniques as basis of many classification applications both theoretically and practically. Before the classifier is built, standard text documents are read, remove stop words and punctuations, stemming the words by using Porter Stemming Algorithm and then features are extracted by using Bigram probability based on keywords such as preprocessing step. The experiment is performed on IEEE and ACM standard documents, research documents. This system is determined the kind of document, such as medicine, computer, engineering and agriculture by using Naïve Bayesian Classifier.*

## 1. Introduction

Classification is the process of finding the common properties among different entities and classifying them into classes.

Classification is one of the data mining approaches which are based on supervised learning.

In supervised learning, an induction algorithm is typically presented with a set of training instances where each instance is described by a vector of feature (of attribute) values, and a class label. The value of an attribute will also called a feature.

The objective of classification is to reduce detail and diversity of data and resulting information overwork by grouping similar data.

A classification model can be used to predict the class label of unknown instants.

The major classification approaches consists of decision tree, K-nearest neighbors, Bayesian approaches, neural networks, regression-based methods and Vector based methods.

The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 describes data preprocessing. Section 4 describes proposed system. Section 5 describes experimental results. Finally concludes this paper in Section 6.

## 2. Related work

Many classification techniques have been used for document classification. Nowadays, the most popular text classification algorithms are Decision Tree Induction, Naïve Bayesian classification, Neural Network, K- Nearest Neighbor and so on.

Chade-Meng Tan, Yuan-Fang Wang, Chan-Do Lee presented to enhance text categorization based on bigram extraction algorithm in which documents are classified into ten categories: agriculture, astronomy, biology, computer science, earth science, engineering, mathematics, physics, space science and zoology.

The algorithm preprocesses the training set to collect the occurrence statistics of each unigram and then to calculate its infogain.

Firstly, the algorithm finds the list of unigrams that appear in a significant number of documents and use them as seeds. And then, documents in the training set are preprocessed to retain only the bodies of each document by discarding headers and the likes. In addition, all numbers and punctuations are removed and all words set to lower case. All stop words are removed using a standard stop words list (Salton and McGill, 1983).

Secondly, the algorithm extracts bigrams where at least one of its component unigrams is a seed. And then, selects only the bigrams among those extracted which high occurrences and infogain. This mean that the bigrams that it select are likely to be good discriminators and less likely to be noisy.

Thirdly, the bigrams are given as features to Naïve Bayesian classifier.

Yahoo-Science Corpus and Reuters-21578 are used for the experiments. The experimental results suggest that the bigrams can substantially raise the quality of feature sets, showing increases in the break-even point and F1-measure [1].

Donghui Feng and Eduard Hovy presented the traditional question answering systems perform the following steps: parsing questions, searching for relevant documents, and identifying/generating answers. N-gram language model use as features and compare the performances by using Decision Tree, Naïve Bayesian classification, SVM (Support Vector Machine) and ME (Maximum Entropy) classification methods. Two corpora of people's biographies, infoplease.com and biography.com, are used for the experiments. The experimental results suggest that N-gram features carry more precise information and therefore should work better than simple ones [2].

## 3. Data preprocessing

Pattern recognition and machine learning has also been applied to document classification. Document classification is an important text mining task because with the existence of a tremendous number of documents or on-line documents. It is tedious yet essential to be able to organize such documents into classes to facilitate document retrieval and subsequent analysis. There are typical classification methods that have been used successfully in text classification. These include nearest-neighbor classification, feature selection methods, Bayesian classification, Support Vector machines, and association based classification.

Data preprocessing is an important issue for both data warehousing and data mining. It can include data cleaning, data transformation, and feature extraction.

In this paper, data preprocessing is a two-step process. In the first step, text documents are read and remove stop words and punctuations. And then, stemming the words by using the Porter Stemming Algorithm. In the second step, features are extracted by using Bigram method based on topic keywords.

Feature selection process can be used to remove terms in the training documents that are statistically uncorrelated with the class labels. This will reduce the set of terms to be used in classification, thus improving both efficiency and accuracy. After the feature selection, which remove non-feature terms, the resulting "cleansed" training documents can be used for effective classification [3]. So we need to extract features for classification.

### 3.1 Bigram

An N-gram is an N-character slice of a longer string. For example, the word "TEXT" would be composed of the following N-grams:

bi-grams: -T,TE,EX,XT,T-
tri-grams: -TE,TEX,EXT,XT-,T--
quad-grams: -TEX,TEXT,EXT-,XT--,T---- [4]

Bigram is a special case of N-gram. Bigram are groups of two written letters, two syllables, or two words. Bigram help provide the conditional probability of a word given the preceding word.

When the relation of the conditional probability is applied:

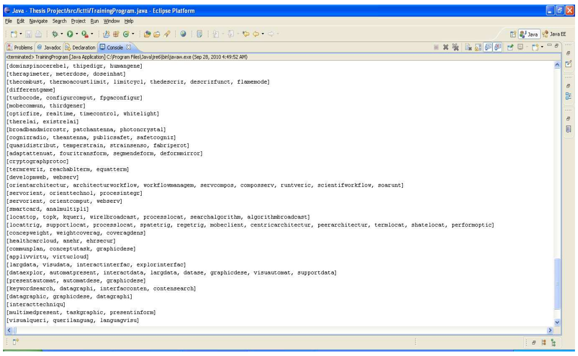
$$\mathbf{P\ (W_n/W_{n-1}) = C(W_{n-1},W_n)/C(W_{n-1})}$$

That is, the probability of a word $W_n$ given the preceding word $W_{n-1}$ is equal to the probability of their bigram or the co-occurrence of the two words $P(W_{n-1},W_n)$ divided by the probability of the preceding word [5].

For Bigram, instead of using the multiplication of conditional probabilities of each word in the Bigram, we only consider the last conditional probability (see below). The reason is that the conditional probability is a strong sign of the pattern's importance. Simply multiplying all the conditional probabilities will decrease the value and require normalization. Realizing that in a set of documents the frequency of each Bigrams is very important information, we combine the conditional probability with the frequency.

$$\mathbf{f\ (W_{n-1},W_n) = P(W_n,W_{n-1}) * f(W_{n-1},W_n)}$$

We consider taking Bigrams in documents as our features. However, Bigram features not closely related to the class label will bring more noise into the system. Therefore, we only take the Bigram within a fixed length window around the topic keywords for features selection [2].


Java - Thesis Project/src/ictti/TrainingProgram.java - Eclipse Platform
File Edit Navigate Search Project Run Window Help
Java
Java EE
Problems
Javadoc
Declaration
Console

```
<terminated> TrainingProgram [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Sep 28, 2010 4:49:52 AM)
[dominspinocerebel, thipedigr, humangene]
[therapimeter, meterdose, doseinhat]
[thecombust, thermoacoustlimit, limitcycl, thedescriz, descrizfunct, flamemode]
[differentgame]
[turbocode, configurcomput, fpgaconfigur]
[mobecommun, thirdgener]
[opticfize, realtime, timecontrol, whitelight]
[therelai, existrelai]
[broadbandmicrostr, patchantenna, photoncrystal]
[cognizradio, theantenna, publicsafet, safetcogniz]
[quasidistribut, temperstrain, strainsenso, fabriperot]
[adaptattenuat, fouritransform, segmendeform, deformmirror]
[cryptographprotoc]
[termrewriz, reachablterm, equatterm]
[developmweb, webserv]
[orientarchitectur, architecturworkflow, workflowmanagem, servcompos, composserv, runtveric, scientifworkflow, soarunt]
[servorient, orienttechnol, procesintegr]
[servorient, orientcomput, webserv]
[smartcard, analmultipli]
[locattop, topk, kqueri, wirelbroadcast, processlocat, searchalgorithm, algorithmbroadcast]
[locattrig, supportlocat, processlocat, spatetrig, regetrig, mobeclient, centricarchitectur, peerarchitectur, termlocat, shatelocat, performoptic]
[concepweight, weightcoverag, coveragdens]
[healthcarcloud, anehr, ehrsecur]
[communplan, conceptutask, graphicdese]
[applivvirtu, virtucloud]
[largdata, visudata, interactinterfac, explorinterfac]
[dataexplor, automatpresent, interactdata, largdata, datase, graphicdese, visuautomat, supportdata]
[presentautomat, automatdese, graphicdese]
[keywordsearch, datagraphi, interfacconten, contensearch]
[datagraphic, graphicdese, datagraphi]
[interacttechniqu]
[multimedpresent, taskgraphic, presentinform]
[visualqueri, querilanguag, languagvisu]
```

**Figure 1: Feature selection by bigram**

### 3.2 Classification

Text mining has become an increasingly popular and essential theme in data mining. Various text mining tasks can be performed on the extracted keywords, tags or semantic information. These include document clustering, classification, information extraction, association analysis and trend analysis.

Text classification is the task of assigning a text document into one or more topic categories or classes [9]. It has a wide range of applications, such as credit approval, customer group identification, medical diagnosis, etc. The problem has been studied extensively by researchers in various fields, such as statistics, machine learning, and neural networks [10].

Nowadays, the most popular text classification algorithms are Decision Tree Induction, Naïve Bayesian Classifier, Neural Network, K-Nearest neighbor and so on.

Classification is a two step process. In the first step, a model is built describing a set of predetermined classes. The set of instances used for model construction is training set. This model is represented as classification rule, decision tree, or mathematical formulae.

In the second step, this model is used for classifying future of unknown objects and to estimate accuracy of the model. If the accuracy is acceptable, use the model to classify instances that class labels are not known. One type of classification patterns used in this paper is the simple Bayesian classifier (also called the Naïve Bayesian Classifier).

### 3.2.1 Naïve Bayesian Classifier

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as probability that a given sample belongs to a particular class. Bayesian classification is based on comparing classification algorithms. It can be described as follows:

Let X be a data sample whose class label is unknown and H be some hypothesis, such as that the data sample X belongs to a specific class C.

For classification problems, it reads to determine P (H/X), the probability that the hypothesis H holds the observed data sample X given.

The probability of a class given by a document P (Class/Document) is dependent on the probability of that class in the document P (Class), has the document P (document) and the likelihood that is given by the class the probability of having the document is P(Document/Class) [6].

**P(Class/Document)=P(Class)*P(Document/Class)/ P(Document)**

A Bayesian classifier is a simple probabilistic classifier based on applying Bayes' theorem. A naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

Depending on the precise nature of the probability model, Naïve Bayesian classifier can be trained very efficiently in a supervised learning setting. Besides good predictive performance, the Naïve Bayesian classifier can also offer a valuable insight into the structure of the training data and effects of the attributes on the class probabilities.

The advantage of the Naïve Bayesian classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification [7].

### 3.2.2 F1- measure (F1- Score)

F1- measure is a measure of a test's accuracy. F1- measure is the harmonic mean of precision and recall. It considers both precision p and recall r of the test to compute the score.

$$F = 2.\frac{precision.recall}{precision + recall}$$

In a statistical classification task, the precision for a class is the number of true positive divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives).

$$Precision = \frac{tp}{tp + fp}$$

The recall for a class is the number of true positive divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives).

$$Recall = \frac{tp}{tp + fn}$$

F1- score reaches its best value at 1 and worst value at 0. The F1- measure is used in the field of information retrieval for measuring search, document classification and query classification performance.

True positives (*tp*) – the number of elements correctly labeled as belonging to the positive class.

False positives (*fp*) – the number of elements incorrectly labeled as belonging to the positive class.

False negative (*fn*) – the number of elements which are not labeled as belonging to the positive class but should have been [8].

## 4. Proposed System

Proposed system has five stages as shown in Figure 2. Training data is used to build the classifier. Testing data is used to test the unknown data set. In the first stage, text documents (text files) are read as training data. In the second stage, training data are preprocessed by stemming the words using the Porter Stemming Algorithm and removing the stop words and punctuation. In the third stage, these training data are used to extract as feature (attribute) by using Bigram method based on topic keywords. The data are partitioned into training data and testing data. We believe the accuracy calculated in this way is more reliable in the classification method. In the fourth stage, these features are used to build the Naïve Bayesian Classifier and evaluate the accuracy of the classifier. And then, system produces the accuracy of the classifier as the result. Finally, unknown data is tested on the testing data in the fifth stage.

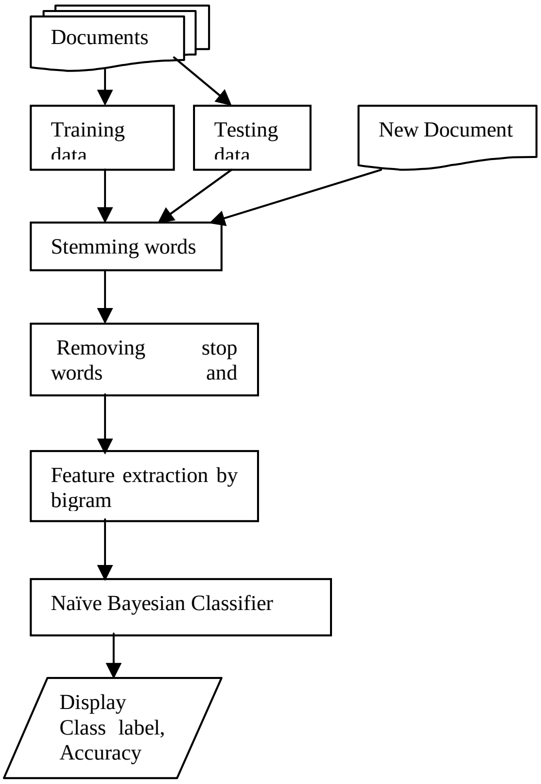


**Figure 2: System flow of classification**

## 5. Experimental Results

The experiments described in this paper to compare Naïve Bayesian Classifier using the accuracy the evolution criteria. The data used for these experiments obtain from IEEE and ACM standard documents and research documents. These data are summarized as data set in Table 1.

| Doc ID | Features | Class |
|---|---|---|
| 1 | breastcanc | medicine |
| 2 | mobecommun, third gener | engineering |
| 3 | developmweb, webserve | computer |
| 4 | ohioclimat, climatsoil | agriculture |
| … | … | … |
| | | |

**Table 1: Dataset used for the system.**

In this system, Accuracy of classifier is measured using the F1-measure of a given datasets into disjoint train and test sets. The classifier is trained on the training dataset and the induced theory is evaluated.

### 5.1 Accuracy

Estimating classifier accuracy is important because it determines to evaluate how accurately a given classifier is. And accuracy also help in the comparison of different classifiers.

Accuracy is calculated by dividing the number of correctly classified documents by the total number of documents, i.e.

$$Accuracy\ (\%) = \frac{correctly\ classified\ documents}{total\ documents} * 100$$

There are 500 documents and 4 classes in this system. These documents are divided into two-third for training data and one-third for testing data. Accuracy is 97%. Table 2 shows the performance evaluation of the system.

| All documents | Training documents | Testing documents | Accuracy |
|---|---|---|---|
| 500 | 333 | 167 | 97% |

**Table 2. Performance evaluation of the system.**

## 6. Conclusion

Classification is the process of finding a set of models that describe and distinguish data classes or concepts. Bayesian classification is based on Bayes theorem. The Naïve Bayesian classification use in focused on the Bayesian formula to calculate the probability of each class given the values of all attributes.

This paper can be applied to predict the class label of unknown sample given the sample data. The relative performance of the Naïve Bayesian classification can serve as an estimate of the conditional independence of attributes. This system involves classification of documents based on features by using Naïve Bayesian classification.

This paper has presented generating of classification from large data sets. This approach demonstrates efficiency and effectiveness in dealing with the classification of documents by using Naïve Bayesian classification.

## 7. References

[1] Chade-Meng Tan, Yuan-Fang Wang and Chan-Do Lee, "The Use of Bigrams to Enhance Text Categorization", Google Inc., 2400 Bayshore Pkwy, Mountain View, CA 94043

[2] Donghui Feng and Eduard Hovy, "Handling Biographical Questions with Implicature", *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processings (HLT/EMNLP),* Vancouver, October 2005, pages 596-603

[3] Jiawei Han and Micheline Kamber, "Data Mining: concepts and techniques, Second Edition",500 Sansome Street, Suite 400, San Francisco, CA 94111,2006

[4] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization", Environmental Research Institute of MIchigan, P.O. Box 134001, Ann Arbor MI 48113-4001

[5] http:// en.wikipedia.org/wiki/Bigram

[6] Markus Forsberg and Kenneth Wilhelmsson, "Automatic Text Classification with Bayesian Learning"

[7] http:// en.wikipedia.org/wiki/Naïve Bayes classifier

[8] http:// en.wikipedia.org/wiki/Precision and recall

[9] Andrew Kachites McCallum, "Multi-Label Text Classification with a Mixture Model Trained by EM", Just Research, Pittsburgh, PA 15213

[10] Hongjun Lu and Hongyan Liu, "Decision Tables: Scalable Classification Exploring RDBMS Capabilities", *Proceedings of the 26th International Conference on Very Large Databases,* Cairo, Egypt, 2000