Scientific Journal Impact Factor (SJIF): 3.518

International Journal of Modern Trends in Engineering and Research

IJMTER

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

# SMS Classification Based on Naïve Bayes Classifier and Semi-supervised Learning

**Miss.Sheetal A. Sable and Prof.P.N. Kalavadekar**

*Department of Computer Engineering,SRES College of Engineering,Kopargaon*

**Abstract—** Short Message Service is one of the most important media of communication due to the rapid increase of mobile users. A hybrid system of SMS classification is used to detect spam or ham, using various algorithms such as Naïve Bayes classifier and Apriori Algorithm. So there is need to perform SMS collection, feature selection, preprocessing, vector creation, filtering process and updating system. Two types of SMS classification exists in current mobile phone and they are enlisted as Black and White. Naïve Bayes is considered as one of the most effectual and significant learning algorithms for data mining and machine learning and also has been treated as a core technique in information retrieval.

**Keywords—**Short Message Service(SMS),Naïve Bayes, Apriori algorithm, ham, spam

## I. INTRODUCTION

Mobile phone has become essential along with the development of wireless communication techniques. Many public institutions and private enterprises utilize the SMSs (Short Message Service) for informing or notifying their customers. Short Message Service has become one of the most important media of communications due to the rapid increase of mobile users. This flood of SMS goes through the problem of spam SMS that are generated by various users. A method is used for building a categorization system is used to integrate association rule mining with the classification problem. However, there is need to perform SMS collection, preprocessing, feature selection, filtering process, vector creation and updating the system. There are two types of SMS classification in the current mobile phones and they are enlisted as Black and White [2]. These techniques are currently available to the number of cell phone operating systems [1].

Naive Bayes is the simplest probabilistic classifiers which is based on Bayes theorem with strong naive independence assumption. This assumption treated each word as a single, mutually exclusive and independent. In the Naïve Bayes classification, all words which are in a given SMS are considered as mutually independent. It is the simplest form of Bayesian network which can be interpreted as conditional independent [8].

## II. LITRATURE SURVEY

There has been numbers of studies on active learning for text classification using probabilistic models, machine learning techniques. The popular techniques for text classifications are Naive Bayes, Support Vector Machine.

### 2.1.Machine Learning Techniques

Automatic text classification has always been considered as an important method to manage and process a vast amount of documents to digital forms that are widespread and continuously increasing. Although machine learning based text classification is good method as far as performance is concerned, it is inefficient for it to handle the very large training corpus. It is good method as far as performance is concerned. Once the system is trained automatically classify the documents. It is very inefficient for it to handle the very large training corpus [3].

### 2.2. Naive Bayes

It is the simplest probabilistic classifiers which are based on Bayes theorem with strong naive independence assumption. This assumption treated each and every word as single, mutually exclusive and independent. Naive Bayes as a probabilistic model is very simple and shows good performance under conditions where the occurring words are independent of each other. With this condition, the Naive Bayes classifier can classify new data only if we count the term frequency occurring in the training samples [8]. It is very simple and shows good performance under conditions where the occurring words are independent of each other. Fast to train and fast to classify. Not sensitive to irrelevant. Handles real and discrete data. Assumes independence of features.

### 2.3.Support Vector Machine

Support Vector Machine is a non-probabilistic classifier in which each document in the data set will be viewed as a point in |v| dimensional space. SVM draws a line in space to separate black points and white points. New incoming documents points will be put in the space. Based on the separating line, we can classify the new incoming messages [5]. It is very simple and shows good performance under conditions where the occurring words are independent of each other. The major drawback is they can be painfully inefficient to train.

## III. SYSTEM DESIGN

### 3.1. System Breakdown Structure

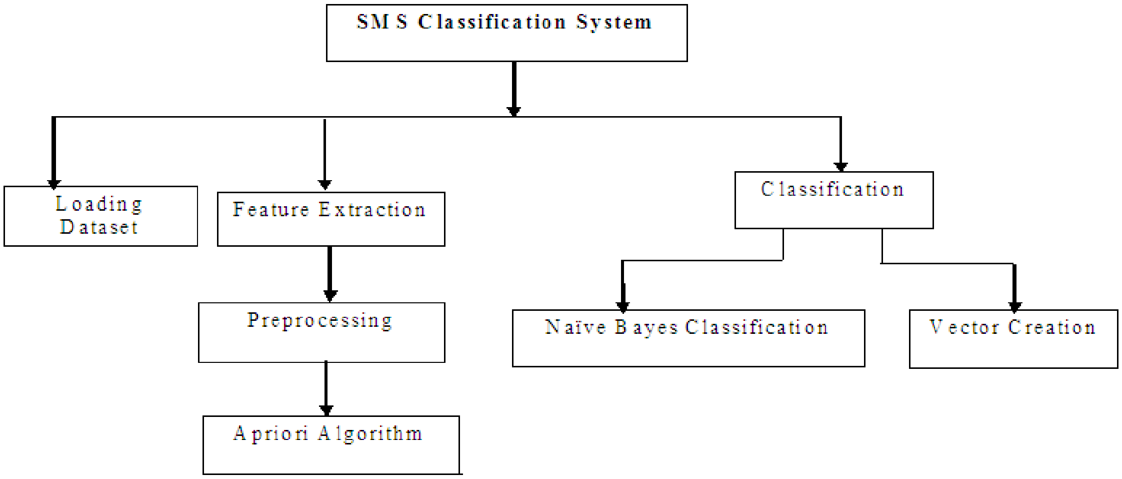

*Figure 1. System Breakdown structure*

### 3.1.1. Load Dataset

This step collects various SMSs from different incoming messages. SMS Spam collection Data Set which consists of SMSs of spam and ham. At the beginning, this database is divided into two subclasses as collection of ham and spam.

### 3.1.2. Feature Extraction

In the traditional Naive Bayes approach, each and every word is considered as an independent word. However, in this approach it is also considered that words are independent to each other, but in modified concept. Additionally, it is treated as the high frequency words as a single and mutually independent also.

$$P(spam|good,very,bad)=P(spam)\ X\ P(good|spam)\ X\ P(very|spam)\ X\ P(bad|spam)$$

### 3.1.2.1. Preprocessing

In preprocessing is used to eliminate the unnecessary words from the SMS. These commonly used words do not play important role in the classification techniques. Therefore they are discarded words.

### 3.1.2.2 Apriori Algorithm

By applying Apriori algorithm,frequent individual items are seperated. However,considering the minimum confidence[4].

- Join Step: Cm is generated by joining Lm-1 with itself
- Prune Step: Any (m-1)- item set that is not frequent cannot be a subset of a frequent m-item set.
- Cm: Candidate itemset of size m
  - Lm: frequent itemset of size m
  - L1= frequent items;
  - for(m= 1; Lm!=Null; m++) do begin
  - Cm+1= candidates generated from Lm;
  - for each transaction t in database do
  - increment the count of all candidates in Cm+1
  - that are contained in t
  - Lm+1= candidates in Cm+1 with minsupport
  - End
  - return ∪m Lm

### 3.1.3. Classification

After building the word occurrence table successfully, run the system to classify a SMS whether the SMS is spam or ham.

### 3.1.3.1 Naive Bayes Algorithm

It is one of the simplest probabilistic classifiers which are based on Bayes theorem with strong naive independence assumption. This assumption treats each and every word as a single, mutually exclusive and independent. The Naive Bayes algorithm is said to be a classification algorithm based on Bayes rule, that assumes all the attributes X1,.,Xn are conditionally and mutually independent given Y. The value of this assumption dramatically simplifies and reduces the complexity and representation of P(X | Y) and the problem of estimating it from the training data.

- Start
- Collect SMS from different incoming messages.
- Assumes all the attributes X1,..,Xn are conditionally and mutually independent given Y.
- Considering the case where X = (X1, X2).
  $P(X|Y) = P(X1,X2|Y)= P(X1|X2,Y)P(X2|Y) = P(X1|Y)P(X2|Y)$
- This can be represented as

$$P(X1|Y)=\pi_{i=1}^{n}=P(Xi|Y)$$

- Calculate the probability that Y can take kth possible value

$$P(Y = yk|X1..Xn) = \frac{P(Y = yk)P(X1..Xn|Y = yk)}{\sum_j P(Y = yj)P(X1..Xn|Y = yj)}$$

- Classify the unknown incoming SMS
  P(ham|good,very,bad) = P(ham) X P(good|ham) X P(very|ham) X P(bad|ham)

### 3.1.3.2. Vector Creation

Vector Creation is very important factor for the Naive Bayes classification system. A dataset is said to be imbalanced, if and only if the classification categories are not approximately represented. This procedure depicts the performance issue of the whole system, this is considered as core part and influence the overall operation.

## IV. CONCLUSION

Automatic text categorization is the task of assigning level of different categorization. This system is going to classify SMS into spam or ham using naive bayes classifier and Apriori algorithm with little bit modification. Although this technique is logic based, but the result is depended with dataset. Supervised machine learning system for handling and organizing spam system and by performing proposed strategy, this SMS spam detection technique have reached accuracy levels that can outperform even the state of the art algorithm.

## REFERENCES

[1] Ishtiaq Ahmed, Donghai Guan, and Tae Choong Chung"SMS Classification Based on Naive Bayes Classifier and Apriori Algorithm Frequent Itemset",International Journal of Machine Learning and Computing, I, Vol. 4, No. 2, April 2014.

[2] Cormack et al., "Spam filtering for short messages", in Proc. The Sixteenth ACM Conference on Conference on Information and Knowledge Management, November 06-10, 2007, Lisbon, Portugal.

[3] M. Ikonomakis, S. Kotsiantis, V. Tampakas,"Text Classification Using Machine Learning Techniques", WSEAS Transactions on Computers, Issue 8, Volume 4, August 2005, pp. 966-974.

[4] P. Madadi, "Text Categorization based on apriori algorithms frequent itemsets", MSc. thesis, School of Computer Science., Howard R. Hughes College of Engineering, University of Nevada, Las Vegas, 2009.

[5] S. Tong and D. Koller,"Support vector machine active learning with applications to text classification", Journal of Machine Learning Research, pp. 45-66, 2001.

[6] A. McCallum and K. Nigam. "A comparison of event models for naive bayes text classification", presented at AAAI-98 Workshop on Learning for Text Categorization, 1998.

[7] S.Weiss, C. Apte, F. Damerau, D. Johnson, F. Oles, T. Goetz, and T. Hampp, "Maximizing text-mining performance",IEEE Intelligent Systems, pp. 6369, 1999.

[8] Z. Cataltepe and E. Aygun. "An improvement of centroid-based classification algorithm for text classification", in Proc. IEEE 23rd International Conference on Data Engineering Workshop, 2007, pp. 952956.

[9] Kyoung-Ju and Deok-Jai Choi, "Mobile Junk Message Filter Reflecting User Preference",KSII transactions on internet and information systems,VOL. 6, NO. 11, Nov 2012.

[10] J. M. G. Hidalgo et al., "Content based SMS spam filtering", in Proc. the 2006 ACM Symposium on Document Engineering, Amsterdam, The Netherlands, October 10-13,
2006.