NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 3, ISSUE 7, July-2016

SMS CLASSIFICATION BASED ON NAIVE BAYES CLASSIFIER AND SEMI-SUPERVISED LEARNING

SHEETAL ASHOKRAO SABLE
Department of Computer Engineering,
SRES College of Engineering, Kopargaon,India

PROF. P.N. KALAVADEKAR
Department of Computer Engineering,
SRES College of Engineering, Kopargaon,India

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 needed to perform SMS collection, feature selection, pre-processing, 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.

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. 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, pre-processing, 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 are 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].

LITERATURE 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.

MACHINE LEARNING TECHNIQUES

Automatic text classification has been considered as an important method to manage and process a vast amount of documents to digital forms that are widespread and continuously increasing. As far as performance is considered. Machine learning text classification is good method, 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 [3].

NAÏVE 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. Naïve Bayes is fast to train and fast to classify. Not sensitive to irrelevant. It handles real and discrete data. Assumes independence of features.

SUPPORT VECTOR MACHINE

SVM 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.

SYSTEM DESIGN

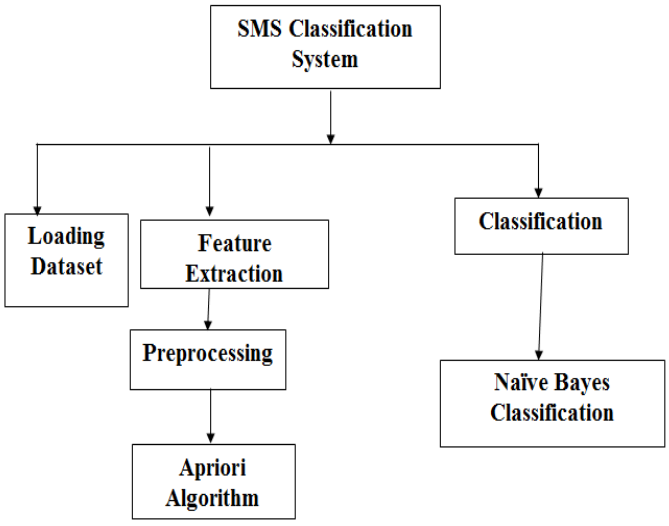


Figure 1: Figure captions should be centred and placed below the figure.

FOLLOWING ARE THE MODULES OF SYSTEM

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.

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.

PREPROCESSING

Pre-processing is used to eliminate the unnecessary words from the SMS. These commonly used words are unnecessary which do not play important role in the classification techniques. Therefore they are known as discarded words.

APRIORI ALGORITHM

By applying Apriori algorithm, frequent individual items are separated. 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 Um Lm

CLASSIFICATION

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

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, 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

MATHEMATICAL MODELLING

A mathematical model is described as a system using mathematical concepts and language. The process used for developing a mathematical model is referred as mathematical modeling.

- **NP-HARD PROBLEMS**
  NP-hard (Non-Deterministic polynomial time hard), in computational complexity theory, is a class of problems that are, informally at least as hard as the hardest problem in NP.

- **NP-COMPLETE PROBLEMS**
  NP-Complete describes decision problems that are the hardest ones in NP in the sense, if there were a polynomially bounded algorithm for an NP-complete problem, then there would be a polynomially bounded algorithm for each problem in NP.

SET THEORY

- Sets
  - Input Set
    - I = I1, I2, I3
    - Where,
    - I1=Username
    - I2=Password
    - I3=SMS Dataset
  - Process Set
    - P= P1, P2, P3
    - Where,
    - P1=Preprocessing
    - P2=Apriori Algorithm
    - P3=Naive Byes and Vector creation
  - Intermediate Output Set

- IO= IO1, IO2, IO3
- Where,
- IO1=Probability of SMS
- IO2=Occurrence of word
- IO3=Classified SMS

- Final Output Set
  - O= O1, O2
  - Where,
  - O1=Spam SMS
  - O2=Ham SMS

VENN DIAGRAM

Venn diagram shows the relation between different inputs, rules, processes and output, it gives the mapping of inputs, processes and outputs. It maps inputs with processes and processes maps to output.

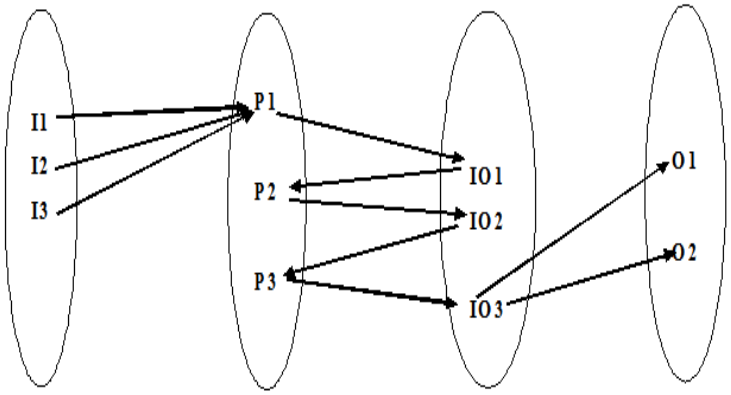


Figure 2. Venn diagram

PROCESS STATE DIAGRAM

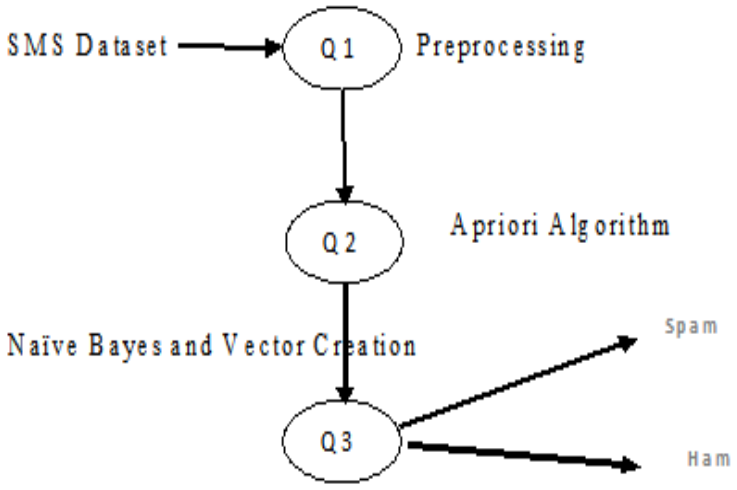
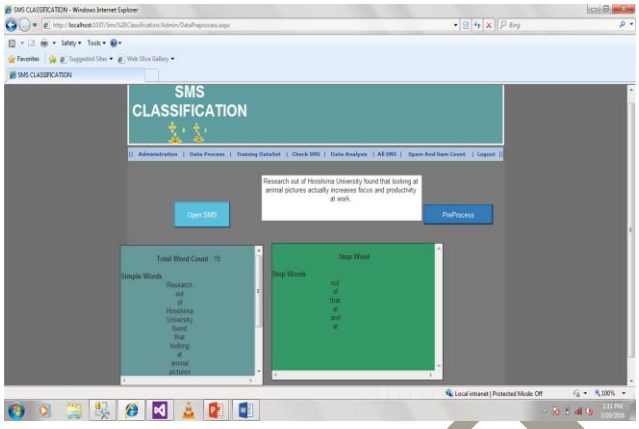


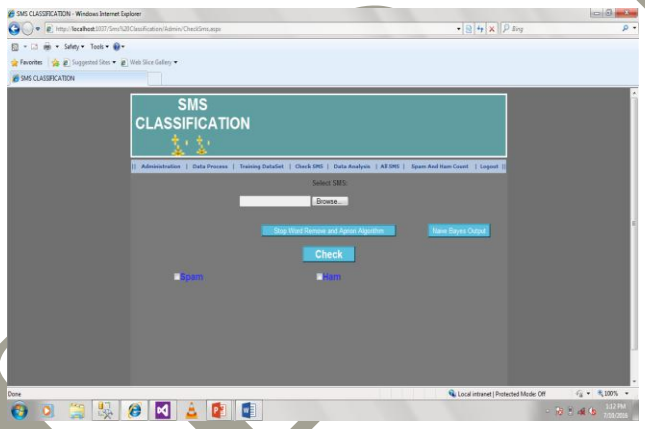
Figure 3. Process State Diagram

IMPLEMENTATION DETAILS

PRE-PROCESSING

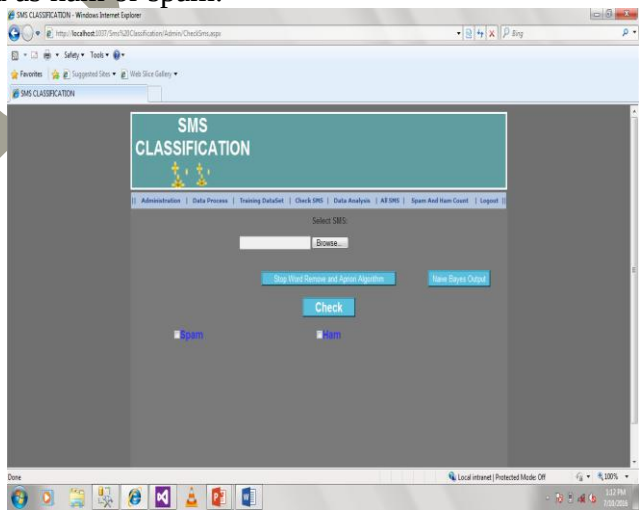


APRIORI ALGORITHM



NAÏVE BAYES

Final output displayed as ham or spam.



RESULT ANALYSIS

RESULT ANALYSIS ON TRAINED DATASET

Standard dataset is collected i.e. from UCI repository. Total Number of SMS tested on trained dataset are 400. Out of which 200 are ham and 200 are spam.

Table 1. Result analysis on trained dataset

No. of Test SMS	Type of SMS		No. of Spam/Ham Detection by system	Implemented System Accuracy
	Ham	Spam		(%)
1 - 100	50	50	100	100
101-200	50	50	100	100
201-300	50	50	100	100
301-400	50	50	100	100

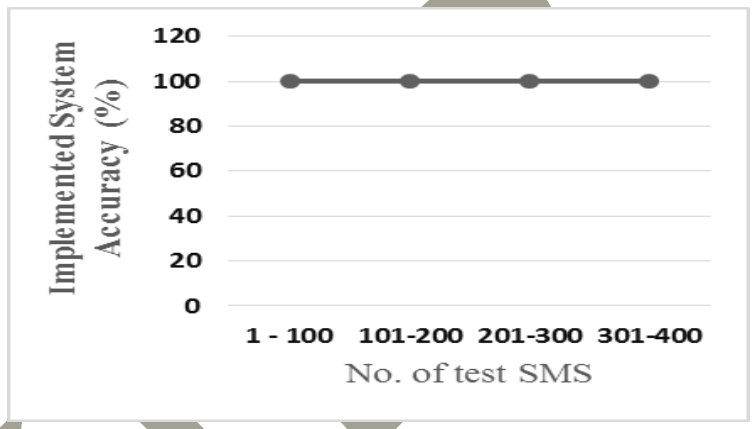


Figure 4. Result Analysis on Trained Data set

RESULT ANALYSIS ON NON-TRAINED DATASET

Total Number of SMS tested on Non-Trained dataset are 200.Out of which 100 are ham and 100 are spam.

Table 2. Result analysis on non-trained dataset

No. of Test SMS	Type of SMS		No. of Spam/Ham Detection by system		Implemented System Accuracy
	Spam	Ham	Spam	Ham	(%)
1-50	5	5	17	17	68
51-100	4	6	15	18	66
101-150	3	7	14	16	60
151 - 200	3	7	13	15	56

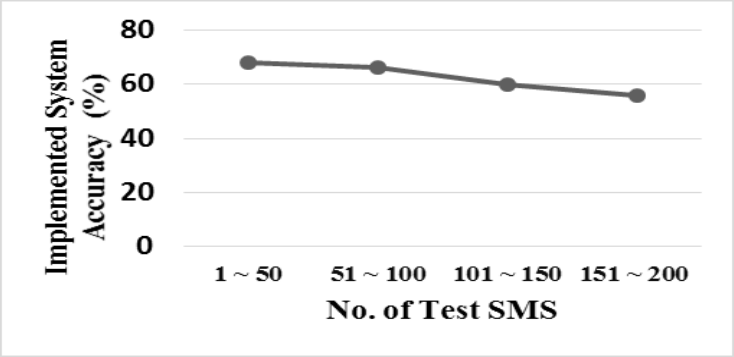


Figure 5. Result Analysis on Non-Trained Data set

PERFORMANCE MEASURE
PERFORMANCE MEASURES USED

To evaluate the performance of classifier commonly used measures are accuracy, precision and recall. Accuracy is given by:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

Precision and Recall is given by:

$$Precision = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

- where TN(true negative)=Ham predicted as ham
- TP (true positive)=Spam predicted as spam
- FP (false positive)=Message not detected as spam
- FN (false negative)= Message not detected as Ham

Accuracy treats the 2 type of errors equally, that is, legitimate classified as spam and spam classified as ham but in SMS classification these 2 errors are not of same importance. A false positive is more costly than false negative. Therefore, accuracy cannot be used here to measure the performance.

- Cost can be computed as: $\text{Cost} = \frac{FP}{TP+FN}$
- The value of cost indicates the percentage of misclassified legitimate SMS from the total spam SMS.
- Thus, its value should be less.
- In other words, lower value of cost depicts better performance.

Table 3. Performance Measure on non- trained dataset

No. of Test SMS	Precision	Recall
1 ~ 50	0.68	0.68
51 ~ 100	0.6	0.68
101 ~ 150	0.56	0.6
151 ~ 200	0.52	0.56

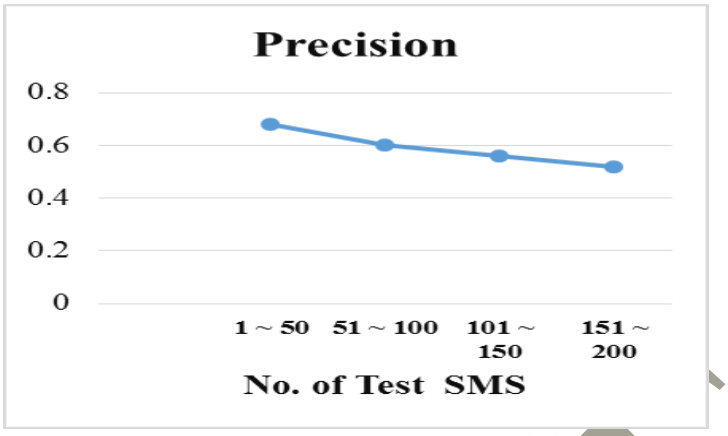


Figure 6. Precision on non- trained dataset

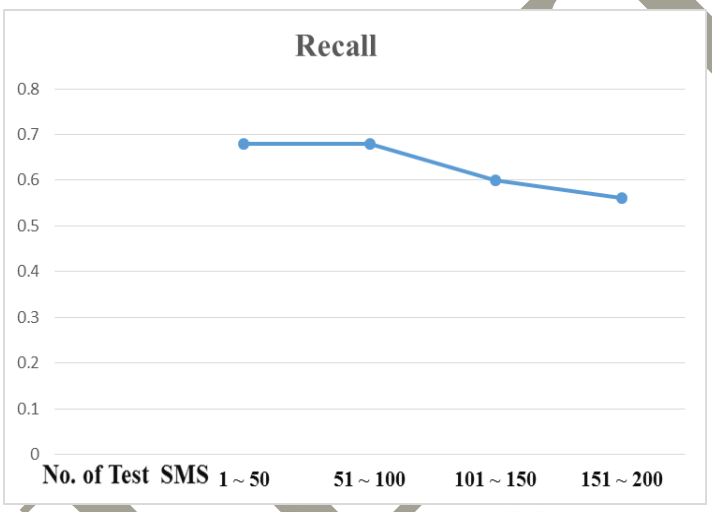


Figure 7. Recall on non- trained dataset

CONCLUSION

This system classifies SMS into spam or ham using Naive Bayes classifier and Apriori algorithm. Although this technique considered as logic based, but the result is depended with dataset. In this, the first module developed is stop word separation, then Apriori and naïve Bayes is applied and final result is generated as spam and ham. The main aim was to detect ham and spam and is detected correctly, this increases overall accuracy. Future extension is when experiment is done on different dataset, exploiting different minimum support produces different result.Hence choosing appropriate minimum support for different corpora would be future interesting research.

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. Mc Callum 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.*