# "Classification of Spam Categorization on Hindi Documents using Bayesian Classifier"

Mr.Ishaan Tamhankar, Dr.Ashysh Chaturvedi
[1]Ph.D Scholar, [2]Department of Computer Science
Calorx Teachers University-Ahmadabad

***Abstract***

*In the current e-world, mostly all the transactions and the business are taking place through e-mails. Now a day, e-mail has become a powerful tool for communication as it saves a lot of time, paper and cost. But, due to social networks sites and advertiser most of the e-mails are containing unwanted information i.e. called* ***spam****. The spam e-mails may contain text of any languages.[3] On the web there are some documents that contain Indian language which may be a spam e-mail. As there are various languages available in India it is a challenging task to identify the spam e-mail due to its linguistic variance and language barriers. As I have reviewed so many research papers on* ***E-mail Spam Categorization****, I found that there are so many classifiers available for all the Indian Language, but there is no document classifier available for* ***Hindi*** *language. So in my research I am going to focus on document classifier for* ***Hindi Spam E-Mail Categorization****.*

**Keywords** - *Hindi Language, Naïve Bayes (NB), Document Categorization, Support Vector Machines (SVM) and K-NN (K – Nearest Neighbors).*

## I. INTRODUCTION

Due to intensive use of Internet, email has become one of the fastest and most economical mode of communication. By this way an Internet user can easily transfer any information from one place to another place through the e-mail in a fraction of seconds. However, with the increase of email users day-by-day it resulted into more increase of spam e-mails during the last few years. E-mail spam is also known as **Junk E-mail** or **Unsolicited Bulk E-mail (UBE).**

There are some predefined categories like **Sports, Health, Entertainment, Business, Astrology, Education, Bank and Spiritual** etc., on which the spam e-mails are sent day-by-day.

The main objective of this system is to enhance the performance of Information Retrieval (IR) and other Natural Language Processing (NLP) applications such as **Library System, Mail Classification, Spam Filtering and Sentiment Analysis** etc., for Hindi language.

India has 23 official languages, with Hindi is chief among them. Hindi uses **Devnagri** script. Hindi is generally spoken in each and every state and even out of India. It is the **4th** most widely spoken language in the world of today. Approximately 310 million peoples are speaking Hindi language in India[5].

### *A. What is a Spam Filter?*

Spam filtering is a procedure of classifying and organizing e-mails based on pre-defined criteria. Often, spam filtering is an automatic procedure where incoming e-mails or messages are classified. This classification can be applied for both incoming and outgoing e-mails[11].

### *B. Naïve Bayes Classifier*

In 1998, the Naïve Bayes classifier was proposed for spam recognition. This technique can be used to classify spam e-mail; word probabilities play the main role here. Naïve Bayes classifier is a collection of classification algorithms based on "Bayes' Theorem". This technique usually uses a set of words to categorize an e-mail as a spam or not.

### *Naïve Bayes classifier works as follows*

They compare the words and / or images that are used in e-mail that can be spam or non-spam and after the classic Bayes' formula a probability is used and probability of an e-mail to be filtered as either a spam e-mail or non-spam e-mail is calculated[8].

### *Using Bayes Theorem for Spam filtering*

Naïve Bayes spam filtering is one of the oldest statistical techniques to filter out a spam e-mail. To implement the Bayesian Filtering for classifying spam e-mails from given set of e-mails [4]. E-mails are majorly classified as spam and non-spam emails. Non-spam emails are also called as ham mails.
The formula used for classifying them is as follows:

$$P(Sp \mid Wd) = \frac{P(Wd \mid Sp) * P(Sp)}{P(Wd \mid Sp) * P(Sp) + P(Wd \mid Ha) * P(Ha)}$$

Here,
P(Sp | Wd) is the probability that a given email is spam given the occurrence of word Wd.
P(Wd | Sp) is the probability that a spam email consists a particular word Wd.
P(Sp) is the probability that an email contains spam contents.
P(Wd | Ha) is the probability that a ham email contains a particular word Wd.

P(Ha) is the probability that an email is not a spam email[12].

### *C. Document Categorization*

Document Categorization is an important task in Information Science and Library Science. The document categorization is classified into certain categories like **Sports, Health, Entertainment, Spiritual, Business, Astrology, Education, and Bank.**

### *D. Proposed Model*

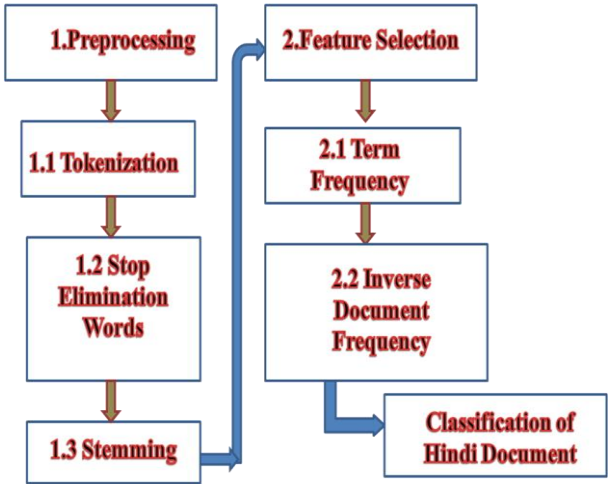


**Table1. Existing Work**

| Sr. No | Author(Year) | Classification Approach | Feature Selection | Data Source | Language | Result/Accuracy |
|---|---|---|---|---|---|---|
| 1 | 1999<br>Support Vector Machines for Spam Categorization | SVM<br>Ripper<br>Rocchio<br>Boosting Decision Trees. | TF-IDF | 1) data set where the number of features were constrained to the 1000 best features<br>2) one data set where the number of features were constrained to the 1000 best features | Chinese | Data sets, boosting trees and SVM had acceptable test performance in terms of accuracy and speed. However, SVM had significantly less training time |
| 2 | 2000<br>Combining Text and Heuristics for Cost – Sensitive Spam Filtering | Druker, Boosting, SVM, Ripper and Rocchio<br>Heuristics for UCE (Unsolicited Commercial E-mail) classification | It uses 4 algorithms:<br>1. NB<br>2. C4.5<br>3. PART<br>4. K-NN | An E-mail message collection containing 4601 message, being 1813 (39%) marked as UCE<br>It uses WEKA 3.0.1 version tool. | English | Best algorithms are C4.5 and PART |
| 3 | 2001<br>Stacking Classifiers for Anti – Spam Filtering of E-Mail | NB, Memory - Based Classifier, Ling - Spam and K-NN | Stacked Generalization | The corpus consists of 2412 linguist messages and 481 spam messages | English | It gives better result with cross-validation stacking |
| 4 | 2002<br>Spam Categorization Through Support Vector Machines | SVM<br>Boosting<br>Ripper | TF-IDF | AT&T staff member and consists of 850 messages that he considered spam and 2150 messages that were non-spam. | English | Boosting has a lower error rate but the dispersion of errors is better using SVM's. |
| 5 | 2004<br>Adversarial Classification | Cost-Sensitive Learning<br>Game Theory<br>NB<br>Spam Detection<br>Integer Linear Programming | Naive Bayes were initially quite successful<br>Adversarial classifier system<br>It Uses 3 scenario:<br>1. Add Word<br>2. Add Length<br>3. Synonym | **Ling-Spam**: On linguistic mailing list there are 2412 non-spam messages and 481 spam ones. It's around 16.6% spam.<br><br>**Email-Data**: Consists of texts from 1431 emails, with 642 non-spam message. | English | False Positive and False Negative for Naive Bayes and the adversary-aware classifier on the Ling-Spam dataset.<br><br>The total number of positives in this dataset is 481, and the total numbers of negatives are 2412. |
| 6 | 2007<br>Spam Detection Using Clustering, Random Forest and Active Learning | Random Forest<br>NB<br>SVM<br>K-NN | Active learning for refining<br>Clustering allows for efficient labelling<br>It uses Partitioning Around Medoids(PAM) algorithm | 9,535 messages of university used for training pool | English | RF – 95.2%<br>NB – 66.7%<br>SVM – 66.7%<br>K-NN – 66.7% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | 2011<br>Machine Learning Methods For Spam E-Mail Classification | NB<br>K-NN<br>ANNs<br>SVMs<br>Artificial immune system and Rough Set | SpamAssassin is used for experiment | It was containing of 824 spam message for testing set | English | NB – 99.46%<br>SVM – 96.90%<br>K-NN – 96.20%<br>ANN – 96.83%<br>AIS – 96.23%<br>RS – 97.42% |
| 8 | 2012<br>Comparative Study on E-Mail Spam Classifier Using Data Mining Techniques | C4.5, C-RT & CS-CRT, ID3, K-NN, LDA (Linear Discriminate Analysis), LR-TRIRLS (Log Regression – Logistic Regression with Truncated Regularized), Multilayer Perception, Naive Bayes, Random Forest Tree and SVM | Here 4 algorithms are used:<br>1. Fisher Filtering<br>2. ReliefF<br>3. STEPDISC (Stepwise Discriminate Analysis)<br>4. Runs Filtering<br>The CART method used under TANAGRA | Tools used TANAGRA data mining tool.<br>In HP Lab – The dataset contains 4601 instances and 58 attributes (57 continuous input attributes and 1 nominal class label target attribute) | English | Random Forest Tree is considered as a best classifier, as it produced 99% accuracy through fisher filtering feature selection. |
| 9 | 2016<br>Proposed Efficient Algorithm to Filter Spam Using Machine Learning Techniques | C4.5 Decision Tree<br>MLP (Multi – Layer Perception)<br>Naive Bayes | To extract vector features of an email, the following methods are used:<br>1. Email header review<br>2. Keyword review<br>3. Black list and White list | The primary data set included 750 valid emails as well as 750 spams.<br><br>Used decision tree, MLP and Naive Bayes by using WEKA tool. | English | NB – 98.6%<br>J48 – 96.6%<br>MLP – 99.3%<br><br>MLP is better than any other algorithms. |
| 10 | 2017<br>Classification of Gujarati Documents Using Naïve Bayes Classifier | NB Statistical Machine Learning Algorithm<br>Decision Tree<br>Neural Network<br>SVM<br>K-NN<br>Rocchio – Style | Used K-fold cross validation to evaluate the performance of NB NB classifier and TF-IDF one used as feature selection | Implementing on 6 categories are:<br>Sports, Health, Entertainment, Business, Astrology and Spiritual<br>280 web documents were collected for each category from various Gujarati News websites.<br>Used 1680 documents including six different categories. | Gujarati | NB classifier without feature selection is 75.74%<br>NB classifier using feature selection 88.96% |

### *E. Methodology*

### *1. Preprocessing*

Main objective of pre-processing phase in document classification is to enhance the influence between word and category of document. It is important step to discard the most insignificant and irrelevant words to improve the quality of document. Steps of pre-processing for document classification as follows:

#### *a. Tokenization*

It is a process to divide texts into number of individual tokens to reduce the unnecessary contents from the document. JAVA utility package and space delimiter were used to done this process. All special characters and punctuation mark have also been removed in this step

#### *b. Stop Word Elimination*

Till now, there is no unique stop words list is available for Indian Hindi language. With the help of linguistic experts and by manual inspection, we have manually constructed a list of stop words. This stop words list is only domain specific that includes sports, entertainment, health, business, spiritual and astrology.

#### *c. Stemming*

For the Hindi language, there is no automation tool is available to create stemmed words list from dataset or corpus. We cam used hand crafted Hindi suffix list in order create a list of stemmed words.

### *2. Feature Selection*

It is the process of selecting most relevant key words from the document based on its frequency and contribution (weight) in the document. In this research, we have used TF-IDF feature selection technique. TF-IDF (Term Frequency-Inverse Document Frequency) weight is a statistical measure which is used to evaluate; how particular word is important for the document from collected dataset or a corpus. Computing functions of TF and IDF are as follows

#### *a. Term Frequency*

Which measure; how frequently a word occurred in a particular document. Frequency of the word is also based on length of the document. Long document may contain more occurrences of the word than short document.

TF calculation all terms to be considered equal importance. TF could be calculated using following formula:

**TF (term) = occurrence of particular term in document/Total numbers of term in Document**

#### *b. Inverse Document Frequency*

Which measure; how particular term is important for the document. IDF could be calculated as logarithm of number of documents in whole corpus

divided by number of document contained particular term. IDF could be calculated using following formula:

**IDF (term) =log(total number of documents in whole corpus/number of document contain a term)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| आ | इ | ई | नी | वान | ता | मान | पा | इया | वासी |
| तर | आनी | कार | आना | ईळा | ओं | ड़ा | आहट | गर | याँ |

## II. LITERATURE REVIEW

1. Harris Drucker, Senior Member, IEEE, Donghui Wu, Student Member, IEEE, and Vladimir N. Vapnik [1999] in their paper 'Support Vector Machines for Spam Categorization'used Classification Approch SVM and Boosting Decision tree.in their Research Paper Focus on Chinese language.Paper Conclude Data Sets,boosting trees and SVM had accepted Performance in terms of Accuracy and speed .

2. Jos4 M. Gomez Hidalgo Manuel Mafia Lopez [2000]," Combining Text and Heuristics for Cost-Sensitive Spam Filtering" in their Paper uses NB,Par,C4.5 and K-NN algoritham for English Langauge and conclude Best Algoritham are C4.5 and Part.

3. Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis,Constantine D. Spyropoulos, and Panagiotis Stamatopoulos [2001] paper "Stacking Classifiers for Anti – Spam Filtering of E-Mail" in research paper aloritham used NB,Ling-Spam and K-NN used 2412 linguist message and 481 spam message data set and give better result with cross-validation stacking.

4. V.DavidSánchez[2004]," Advanced support vector machines and kernel methods"in article used alogritham Ripper and Boosting in which dataset AT&T Staff member and consists of 850 message that consider spam and 2150 messagesd that were non-spam as a result proved Boosting has low error rate but the dispersion of Error is better using SVM's.

5. Nilesh Dalvi Pedro Domingos Mausam Sumit Sanghai Deepak Verma[2004] "Adversarial Classification" had implemented adversarial classification system for spam filtering domain ACS use 3 Scenario : 1) Add Word. 2) Add Length 3) Synonym.Naive Bayes were initially quite Successful. False Positive and False Negative for Naïve Byas and adversary –aware classifier on Ling-Spam Data set.

6. Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis[2006]," Automatic Text Categorization in Terms of Genre and Author" in their paper presented an approach to text categorization in terms of genre and author for Modern Greek. In contrast to previous stylometric approaches, we attempt to take full advantage of existing natural language processing (NLP) tools

7. Dave DeBarr, Harry Wechsler, PhD [2007] "Spam Detection using Clustering Random Forest and Active Learing" in research paper focused on efficient construction of effective models for spam detection.Clustering messages allows for efficient labeling of a representative sample of messages for learning a spam detectionmodel using a Random Forest for classification and active learning for refining the classification model.Data set 9535 messages of university used for Trainning pool as a result RF-95.2%,NB-66.7%,SVM -66.7% and K-NN-66.7%.

8. Ismaila Idris [2011]," E-mail Spam Classification With Artificial Neural Network and Negative Selection Algorithm" in their Research Paper apply neural network and spam model based on Negative selection algorithm for solving complex problems in spam detection. It consisting 824 spam message for testing set, as a result NB-97.46% ,SVM-96.90%,K-NN-96.20%,ANN-96.83%,AIS-96.23% and RS-97.42%

9. R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, Member, IAENG," Comparative Study on Email Spam Classifier using Data Mining Techniques" in their research used 4 Alogoritham 1) Fisher Filtering 2) ReliefF 3) STEPDISC and used tools TANAGRA as a result Random Forest Tree is consider as best classifier as it produced 99% accuracy through fisher filtering feature selection.

10. Ali Shafigh Aski,Navid Khalilzadeh Souratib [2016] " Proposed Efficient Algoritham to Fileter spam using Machine Learing Techniques" used C4.5 Decision Tree , MLP and Naïve Byas for Extract vector features of Email for Following methods 1) Email Header Review 2) Keyword Reviwe 3) Blacklist and White list.Paper give Result NB-98.6% J48 – 96.6% and MLP -99.3%

11.Diab M.DiabKhalil M.El Hindi [2016]," Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification" in paper using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification.

12. Rajnish M. Rakholia and Jatinderkumar R. Saini [2017]," Classification of Gujarati Documents using Naïve Bayes Classifier" in their research Paper used K-fold validation to evaluate the performance of NB.NB classifier and TF-IDF used as feature selection and implementing in various categories like Sports, Entertainment, Business, Astrology and

Spiritual. NB classifier without feature selection is 75.74%. NB classifier using feature selection 88.96%.

13. Nidhi, et. al. [6] presented for the first time domain based classification of Punjabi text documents using ontology and Hybrid approach (combination of Naïve Bayes and Ontology based classification). They chose Sport domain for creating ontology manually. Their results shows that these approaches provide better results compared to standard algorithms such as Naïve Bayes classifier (NB) and Centroid classifier

14. Kavi Narayana Murthy [7] proposed automatic text classification for Telugu news articles using Naïve Bayes(NB) classifier. The four major categories defined include Politics, Sports, Business and Cinema. The performance of NB is computed in terms of precision, recall and F-measure.The author's technique does not use stop word removal,stemming and morphological analysis.The review on existing literature reveals that not much workhas been carried out for the text classification of Indian regional languages. Some of the supervised learning methods applied include K-Nearest Neighbor (KNN), Modified K-Nearest Neighbor (MKNN), Centroid algorithm, Naïve Bayes(NB), and Support Vector Machine (SVM) on languages like Bangla, Marathi, Tamil, Telugu, Punjabi and Urdu.Among the classification techniques MKNN, KNN, Naïve Bayes, Centroid and one of the clustering techniques i.e.LINGO algorithm applied on Marathi language. These techniques exclude stop word removal and morphological analysis which would have given better results.

## III. CONCLUSION AND FUTURE WORK

This work has been carried out to Hindi document classification using Naïve Bayes classifier. We have also discussed the results of classifier for multi-category Hindi documents. We can achieved good accuracy by which is more influence and related to the particular domain specific category. NB classifier consider each word as an independent word in document and needs training to implement.

In future we will apply Filteration for Hindi document classification and extend this work by adding new category in Ontology which can be used in other research in area of Natural Language Processing and Mining

## REFERENCES


[1] Lin SH, Chen M C, Ho JM, Huang YM. ACIRD: Intelligent Internet document organization and retrieval. IEEE Transactions on Knowledge and Data Engineering. 2002; 14(3):599–614.https://doi.org/10.1109/ TKDE.2002.1000345

[2] Lee LH, Isa D. automatically computed document dependent weighting factor facility for Naïve Bayes classification. Expert Systems with Applications, 2010; 37(12):8471–8. https://doi.org/10.1016/j.eswa.2010.05.030

[3] Zhang H. The Optimality of Naive Bayes. Barr V, Markov Z, editors. FLAIRS Conference; AAAI Press; 2004.

[4] Patil JJ, Bogiri N. Automatic text categorization Marathi documents. International Journal of Advance Research in Computer Science and Management Studies. 2015; 3(3):280–7. https://doi.org/10.1109/icesa.2015.7503438

[5] Patil M, Game P. Comparison of Marathi text classifiers. ACEEE International Journal on Information Technology. 2014; 4(1):11–22.

[6] mandal ak, sen r. supervised learning method for bangla web Document Categorization. International Journal of Artificial Intelligence and Applications. 2014; 5(5):93–105. https://doi.org/10.5121/ijaia.2014.5508

[7] Murthy VG, Vardhan BV, Sarangam K, Reddy PVP. A comparative study on term weighting methods for automated Telugu text categorization with effective classifiers. International Journal of Data Mining and Knowledge Management Process. 2013; 3(6):95. https://doi. org/10.5121/ijdkp.2013.3606

[8] Swamy MN, Hanumanthappa M. Indian language text representation and categorization using supervised learning algorithm. International Journal of Data Mining Techniques and Applications. 2013; 2:251–7.

[9] Naseeb N, Gupta V. Domain based classification of punjabi text documents using ontology and hybrid based approach. Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing COLING; 2012. p. 109–122.

[10] Rajan K, Ramalingam V, Ganesan M, Palanivel S, Palaniappan B. Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Systems with Applications. 2009, 36(8):10914–8. https://doi.org/10.1016/j.eswa.2009.02.010

[11] Raghuveer K, Murthy KN. Text categorization in Indian languages using machine learning approaches. IICAI; 2007. p. 1864–83.

[12] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2002; 10:79–86.

[13] Rogati M, Yang Y. High-performing feature selection for text classification. Proceedings of the 11th International Conference on Information and Knowledge Management; 2002. p. 659–61. https://doi.org/10.1145/584792.584911

[14] Forman G. An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research. 2003; 3:1289–305.

[15] Tan S, Zhang J. An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications. 2008; 34(4):2622–9. https://doi.org/10.1016/j. eswa.2007.05.028

[16] Prabowo R, Thelwall M. Sentiment analysis: A combined approach. Journal of Informetrics. 2009; 3(2):143–57. https://doi.org/10.1016/j.joi.2009.01.003

[17] Alsaleem S. Automated Arabic text categorization using SVM and NB. International Arab Journal of e-Technology. 2011; 2(2):124–8.

[18] El Kourdi M, Bensaid A, Rachidi TE. Automatic Arabic document categorization based on the Naïve Bayes algorithm. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Association for Computational Linguistics; 2004. p. 51–8. https://doi. org/10.3115/1621804.1621819

[19] Hadni M, Lachkar A, Ouatik SA. A new and efficient stemming technique for Arabic text categorization. 2012 International Conference on Multimedia Computing and Systems (ICMCS); 2012. p. 791–6. https://doi.org/10.1109/ ICMCS.2012.6320308

[20] Harrag F, El-Qawasmah E, Al-Salman AMS. Stemming as a feature reduction technique for Arabic text categorization. 2011 10th International Symposium on Programming and Systems (ISPS); 2011. p. 128–33.

[21] Halder T, Karforma S, Mandal R. A novel data hiding approach by pixel-value-difference steganography and optimal adjustment to secure e-governance documents.

Indian Journal of Science and Technology. 2015 Jul; 8(16):1–7. https://doi.org/10.17485/ijst/2015/v8i16/51269
[22] Prakash KB. Mining issues in traditional Indian web documents. Indian Journal of Science and Technology. 2015 Nov; 8(32):1–11.
[23] Antipov KV, Vinokur AI, Simakov SP, Isakov YV, Kazakova AY. Digitization of Russian parish registers of the 18-20th centuries as the contribution to the cultural foundation of historical documents. Indian Journal of Science and Technology. 2015 Dec; 8(10):1–10. https://doi.org/10.17485/ijst/2015/v8is(10)/87462
[24] Posonia AM, Jyothi VL. Context-based classification of XML documents in feature clustering. Indian Journal of Science and Technology. 2014 Jan; 7(9):1–4.
[25] Karthika S, Sairam N. A naïve bayesian classifier for educational qualification. Indian Journal of Science and Technology. 2015,Jul;8(16):1–5. https://doi.org/10.17485/ijst/2015/v8i16/62055
[26] Sarangi PK, Ahmed P, Ravulakollu KK. Naïve Bayes classifier with LU factorization for recognition of handwritten Odia numerals. Indian Journal of Science and Technology. 2014 Jan; 7(1):1–4.