International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 10 | Oct - 2021 ISSN: 2582-3930

An Implementation on Email Classification on Hindi Language using Bayesian Classifier

Mr.Ishaan Tamhankar[1]
Assistant Professor
V.T Poddar BCA College

Ms. Ritu Bhatiya[2]
Assistant Professor
V.T Poddar BCA College

Abstract:

Spam messages are one of the most serious issues on the Internet today, causing financial harm to businesses and annoyance to individual users. Spam filtering can help with the problem in a number of ways. The classifier-related challenges have been the focus of several spam filtering studies. Machine learning for spam classification is now a significant research topic. The application of various machine learning techniques for categorizing spam messages from e-mail is investigated and identified in this research. Finally, with spam categorization, a comparative study of the algorithms has been presented.

Keywords:

Spam, Email Classification, Machine Learning, Naïve Bayes

Introduction:

Spam, or unwanted commercial or bulk e-mail, has recently become a major issue on the internet. Spam is a waste of time, storage space, and data transfer capacity. Spam e-mail has been on the rise for several years. According to recent data, spam accounts for 40% of all emails, or 15.4 billion each day, costing internet users $355 million every year. Automatic e-mail filtering appears to be the most successful approach for combating spam at the present, and spammers and spam-filtering technologies are competing fiercely. E-mail filtering uses two general approaches: knowledge engineering and machine learning. A set of rules must be provided in the knowledge engineering technique to classify emails as spam or ham. A collection of such rules should be created by either the filter's user or by another authority (for example, the software business that provides a specific rule-based spam-filtering tool). This strategy yields no promising results because the rules must be changed and maintained on a regular basis, which is a waste of time and inconvenient for most users. Machine learning is more efficient than knowledge engineering since it does not necessitate the specification of any rules. Instead, a set of training samples is used, which consists of a collection of pre-classified e-mail messages. The categorization rules are then learned using a specific algorithm from these e-mail communications. Machine learning has been extensively researched, and there are numerous algorithms that may be employed in e-mail filtering. Naive Bayes and artificial Neural Network are a few examples.

Related Work:

There are some research work that apply machine learning methods in e-mail classification, Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali[2] They demonstrated that the naïve Bayes e-mail content classification

could be adapted for layer-3 processing, without the need for reassembly. Suggestions on predetecting e-mail packets on spam control middleboxes to support timely spam detection at receiving e-mail servers were presented. M. N. Marsono, M. W. El-Kharashi, and F. Gebali[1] They presented hardware architecture of na¨ıve Bayes inference engine for spam control using two class e-mail classification. That can classify more 117 millions features per second given a stream of probabilities as inputs. This work can be extended to investigate proactive spam handling schemes on receiving e-mail servers and spam throttling on network gateways. Y. Tang, S. Krasser, Y. He, W. Yang, D. Alperovitch [3] proposed a system that used the SVM for classification purpose, such system extract email sender behavior data based on global sending distribution, analyze them and assign a value of trust to each IP address sending email message, the Experimental results show that the SVM classifier is effective, accurate and much faster than the Random Forests (RF) Classifier. Yoo, S., Yang, Y., Lin, F., and Moon [11] developed personalized email prioritization (PEP) method that specially focus on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of particular user, as well as they developed a supervised classification framework for modeling personal priorities over email messages, and for predicting importance levels for new messages. Guzella, Mota-Santos , J.Q. Uch, and W.M. Caminhas[4] proposed an immune-inspired model, named innate and adaptive artificial immune system (IA-AIS) and applied to the problem of identification of unsolicited bulk e-mail messages (SPAM). It integrates entities analogous to macrophages, B and T lymphocytes, modeling both the innate and the adaptive immune systems. An implementation of the algorithm was capable of identifying more than 99% of legitimate or SPAM messages in particular parameter configurations. It was compared to an optimized version of the naive Bayes classifier, which have been attained extremely high correct classification rates. It has been concluded that IA-AIS has a greater ability to identify SPAM messages, although the identification of legitimate messages is not as high as that of the implemented naive Bayes classifier

Methodology

Naïve Bayes Classifier Working Model:

Hypothesis A opportunities in Visual Event P(A) P(A | B), has a Posterior option.
P (B | A) Opportunities: Evidence opportunities if the hypothesis of probability is true.
P(A) is given an earlier opportunity: hypothesis chances before proof is seen.
With Margin: Evidence Opportunity, P (B) is possible.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

The following model may be used to understand how the Nave Bayes' Classifier works:

Suppose we have a weather dataset and a target variable called "play." Therefore we have to determine whether we would play in accordance with the conditions of weather to use this data set on a certain day. We must take the following steps to tackle this problem:
Turn the data set into frequency tables. Using this Model we are going Classified Spam Emails in

Astrology, Bank, Education, Entertainment, Others, Shopping, Sports in Various Categories.

Spam Email Data Set:

From	To	Subject
services@custcomm.icicibank.com	ishaan.tamhankar06@outlook.com	आपके विज़न बोर्ड 2020 में आपके वित्तीय लक्ष्य
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	धनु राशि में बृहस्पति का पारगमन 2020: आपके चंद्रमा पर प्रभाव
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	मेष और मीन 2020 में मंगल का प्रतिगमन: चंद्रमा के संकेतों पर प्रभाव
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	राहु केतु गोचर 2020: 12 चंद्र राशियों के लिए प्रभाव और भविष्यफल
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	पितृ पक्ष 2020 तिथियां, नियम और पूजा प्रक्रिया
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	ज्योतिष के अनुसार अपना लकी रंग और नंबर जानिए
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	ज्योतिष परामर्श से वित्तीय स्थिति में सुधार के उपाय
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	रिश्ता दुविधा? इसे ज्योतिष परामर्श से हल करें
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	संक्रांति	वर्ष 2020 का सबसे लंबा दिन
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	सभी चंद्र राशियों पर पेनुमब्रल चंद्र ग्रहण का प्रभाव
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	अक्षय तृतीया / अखा तीज 2020 तिथि, महत्व और अनुष्ठान
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	आपकी राशि पर चैत्र नवरात्रि 2020 के प्रभाव
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	विवाह में देरी के कारण और उपचार, देर से विवाह ज्योतिष
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	कल्कि अवतार: भगवान विष्णु का दसवां और अंतिम अवतार
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	ज्योतिष में योनी मिलान का महत्व
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	विवाह के लिए नक्षत्र या तारा मिलान
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	राशि चक्र संकेत के अनुसार पहली तारीख गाइड
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	ग्रिहाय प्रवीण मुहूर्त या तिथियां २०२० में - तीथि, समय और नक्षत्र
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	बिजनेस न्यूमेरोलॉजी - बिजनेस सक्सेस के लिए लकी नंबर
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	व्यापार में सफलता के लिए 5 सबसे शक्तिशाली मंत्र
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	शेयर बाजार की भविष्यवाणी के लिए योगासन	शेयर बाजार ज्योतिष
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	अपनी कुंडली के माध्यम से अपनी संपत्ति की संभावनाओं को जानें
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	मेरी शादीशुदा जिंदगी कैसी होगी? विवाह संबंधी समस्याएं और ज्योतिष
ganesha@gmail.ganeshspeaks.com	ishaan.tamhankar06@outlook.com	गोमेद पहनने के ज्योतिषीय लाभ	हेसोनाइट जेमस्टोन

Steps of Algorithm:

Step-1 Data Pre-processing step
Step-2 Fitting Naive Bayes to the Training set
Step-3 Predicting the test result
Step-4 Test accuracy of the result (Creation of Confusion matrix)
Step-5 Visualizing the test set result.

Data Pre-Processing Step:

```
import  pickle
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer

df=pd.read_csv("emailsclassi.csv")
x = df["Subject"]
y = df["feature"]

# x_train,y_train = x[0:560],y[0:560]
# x_test,y_test = x[560:],y[560:]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)

##Step3: Extract Features
cv = CountVectorizer()
features = cv.fit_transform(x_train)
```

Fitting Naive Bayes classifier to the Training data:

In training data we are now going to equal the Naive Bayes divider. In this regard, we are introducing the sklearn.naive bayes library's MultinomialNB section. We will create a class divider object after introducing the class. Then, in the training data, we measure the separator. Underneath your code:

```
#Fitting Naive Bayes classifier to the training set
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
nb.fit(features,y_train)
```

Output: if you execute the above code, the output is as follows

```
Out[24]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

Predicting Test Results: We will create a y pred vector as in the logistic regression in order to predict the test of set results. Underneath your code:

```
import sys
from time import time
from sklearn.metrics import accuracy_score
# print ("Training time:", round(time()-t0, 3), "s")
t1=time()
y_pred=nb.predict(features)
print ("Prediction time:", round(time()-t1, 3), "s")
print ("Accuracy Score",accuracy_score(y_train,y_pred))
```

OUTPUT:

```
Prediction time: 0.001 s
Accuracy Score 0.9142857142857143
```

Creating the Confusion Matrix:

In order to see the precision of the split, we will build a confusion matrix for our Naive Bayes model now. Underneath your code:

```
from sklearn.metrics import multilabel_confusion_matrix
cm = multilabel_confusion_matrix(y_train, y_pred)
print(cm)
```

We can therefore say that the performance in the model is improved by means of the K-NN algorithm in the above chart, 532 + 17 = 549 correct predictions, and 8 + 3 = 11 incorrect forecasts.

Visualizing the Training set result:

The training results for the model from Naive Bayes will now be visualised. With the exception of the graph name, the code is always the same as the KNN and SVM code. Underneath your code:

```
from sklearn.preprocessing import LabelEncoder
lb=LabelEncoder()
cl=lb.fit_transform(y_train)

plt.scatter(x_train, classifier.predict(cv.transform(x_train)),
c=cl, cmap='winter')
plt.show()
plt.close()
```

OUTPUT :

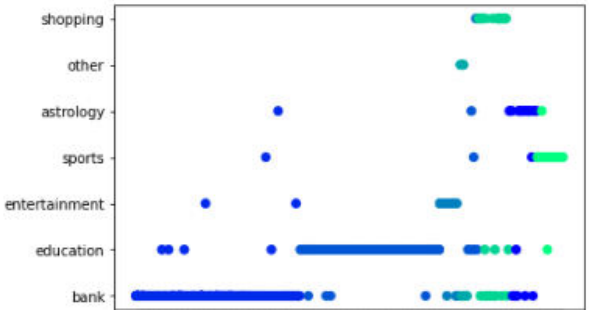

Figure 1. NB Visualizing Spam Email Data

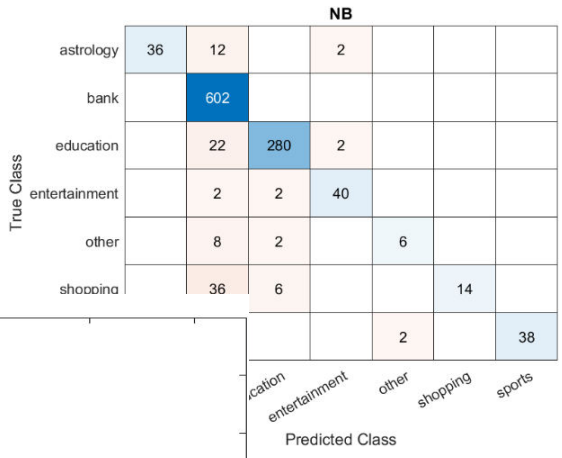

Figure 2. NB Confusion Matrix

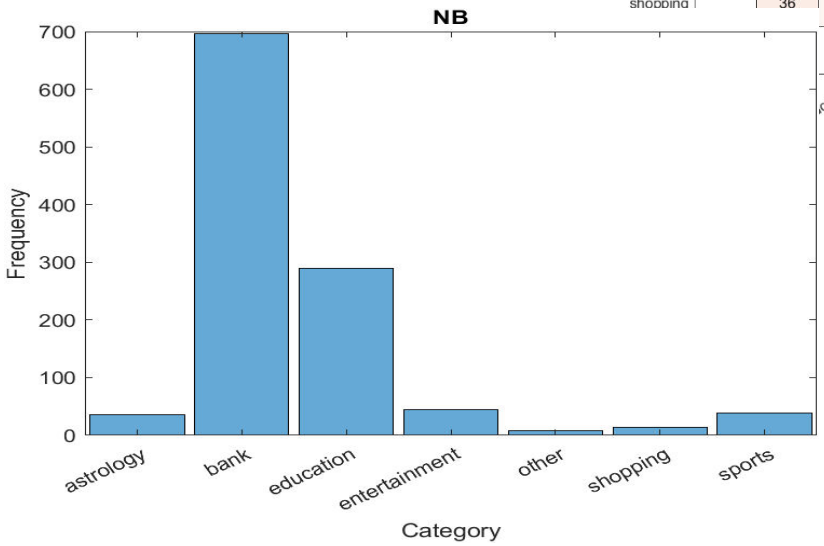

Figure 3. Classified Spam Email for Naïve Bayes

Conclusion:

The results of the data test show that the fundamental goal is achieved and the classification results are achieved. This section uses the NB machine learning classification. Hence in this Implementation Model Achieved **91% Accuracy** for Classified Data Set. The Algorithm for the NB division is yours, as the distance scale must be set. Since distance understanding is not profound, the effect of separation is entirely dependent on the used distance. For this reason, experts need to assess whether the result is working with a set of data, two separate algorithms that produce two completely different outcomes. Since it is often dynamic to recognize results, the application of different grades is reduced.

References:

[1] S.J. Delany, P. Cunningham, and L. Coyle, "An assessment of case-based reasoning for spam filtering", Artificial Intelligence Review Journal, Vol. 24, No. 3-4, 2018, pp. 359-378.

[2] P. Cunningham, N. Nowlan, S.J. Delany, and M. Haahr, "A case-based approach in spam filtering that can track concept drift", In Proceedings: The ICCBR"03 Workshop on Long-lived CBR Systems, Trondheim, Norway, 2016

[3] K. Wei, A naïve Bayes spam filter, Faculty of Computer Science, University of Berkely, 2012.

[4] B. Kamens, Bayesian filtering: Beyond binary classification. Fog Creek Software, Inc., 2010.

[5] M.I. Devi, R. Rajaram, and K. Selvakuberan, "Generating best features for web page classification", Webology, Vol. 5, No. 1, 2008, Article 52.

[6] M. Hartley, D. Isa, V.P. Kallimani, and L.H. Lee, "A domain knowledge preserving in process engineering using self-organizing concept", In Proceedings: ICAIET 06. Sabah, Malaysia: Kota Kinabalu, 2006.

[7] X. Su, A text categorization perspective for ontology mapping, Norway: Department of Computer and Information Science, Norweigian University of Science and Technology, 2002.

[8] E.H. Han, G. Karypis, and V. Kumar, Text categorization using weight adjusted k-nearest neighbour classification, Department of Computer Science and Engineering, Army HPC Research Center, University of Minnesota, 1999.

[9] A. McCallum, and K. Nigam, "A comparison of event models for naïve Bayes text classification", Journal of Machine Learning Research, Vol. 3, 2003, pp. 1265–1287.

[10] S. Chakrabarti, S. Roy, and M.V. Soundalgekar, "Fast and accurate text classification via multiple linear discriminant projection", The VLDB Journal The International Journal on Very Large Data Bases, 2003, pp. 170–185.

[12] I. Rish, An empirical study of the naive Bayes classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. (available online: PDF, PostScript).

[13] M. Mozina, J. Demsar, M. Kattan, and B. Zupan, Nomograms for Visualization of Naive Bayesian Classifier, In Proc. of PKDD-2004, pages 337-348. (available online: PDF), 2004.

[14] O. R. Duda, P. E. Hart and D. G. Stork, Pattern classification (2nd edition), Section 9.6.5, p. 487-489, Wiley, ISBN 0471056693,2000.

[15] J.R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1993. [12] S. Wermter, "Neural network agents for learning semantic text classification", Information Retrieval, Vol. 3, No. 2, 2004, pp. 87-103.

[16] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification", In Proceedings: IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61–67, 1999.

[17] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", In Proceedings: Machine Learning: ECML-98, 10th European Conference on Machine Learning, pp. 137–142, 1998.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update", SIGKDD Explorations, Vol. 11, No. 1, 2009, pp. 10-18

[19] Richard Power. 1999 CSI/FBI computer crime and security survey. Computer Security Journal, Volume XV (2), 1999.