ANALYSIS OF EMAIL INTO SPAM AND HAM USING MULTINOMIAL NAÏVE BAYES

A. N. ARULARASAN[*1], RAVIRAJAN S R[2], DHIVAKARAN K[3], ASHOK KUMAR V[4]

[1]Associate Professor, Department of Artificial Intelligence and Data Science,
Panimalar Engineering College, Chennai. India.
[2]UG Student, Department of Artificial Intelligence and Data Science,
Panimalar Institute of Technology, Chennai. India
[3]UG Student, Department of Artificial Intelligence and Data Science,
Panimalar Institute of Technology, Chennai. India.
[4]UG Student, Department of Artificial Intelligence and Data Science,
Panimalar Institute of Technology, Chennai. India.

Abstract

The most difficult problems in social networks is spam identification. (OSNs). In this study, spam on these sites is found using a supervised technique. The choice of the intended features and the application of the right classifier both had an impact on the precision of supervised techniques. The first element is also handled in a novel way. With the intention of selecting the desired characteristic from a range of features, this technique combines association rule mining and genetic algorithms. The second component, on the other hand, makes use of a lot of well-liked predictors. The efficacy of the proposed feature selection method on the classifiers precision is demonstrated by the evaluation of the proposed method on three datasets. The combined sum of both methodologies versus the fundamental techniques are 87.98% and 95.23%, respectively.

Keywords : Spam, ham, Spam detection, Multinomial Naïve Bayes , Bayes Theorem ,Phishing

1. Introduction

Spam is any unsolicited, bulk transmission of any undesirable digital communication. In addition to social media, text messaging, and phone calls, spam can also be distributive via email. A virus that spreads over the air is referred to as "Spam". The term “spam” began to be used to designate a sizable fraction of unwelcome messages because of a Monty python sketch in which the actors assert that everyone must eat spam, whether they want to or not. Similarly, whether we like it or not, spam emails must regrettably annoy everyone with an email account. Spam filters are used to recognise these emails in order to prevent unsolicited, unpleasant, and virus-infected emails from getting to a user's mailbox. Like other filtration programmes, a spam filter looks for certain criteria before making a conclusion. For instance, whenever a user marks an email as junk, the Bayesian filter notices the trend and instantly transfers any subsequent emails from that sender to the spam folder.

Emails provide secrecy because only the author and the recipient can see the communication. Businesses can share comprehensive information via email by attaching documents like spreadsheets and word reports. Customized email platforms' additional security feature enables businesses to manage the communications. In this instance, there are two kinds of mail:

HAM -Valid mails

SPAM – Unwanted mails

The spam categorization system (also known as the nave Bayes classifier) is developed in this system to distinguish between spam and non-spam. The classification model for email spam is displayed in Figure 1.

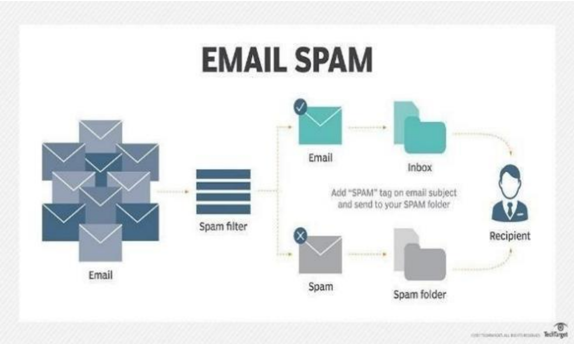
EMAIL SPAM
Email
Spam filter
Email
Inbox
Add "SPAM" tag on email subject and send to your SPAM folder
Recipient
Spam
Spam folder

Fig 1. Classifier diagram

3. Related Works

[1] The authors claimed that the rise in the volume of spam unsolicited bulk e-mails has necessitated the development of reliable anti-spam filters. The article went on to say that automatic spam email screening has been made possible by machine learning techniques. The author then reviewed some of the most well-liked machine learning techniques, including k-NN, ANNs, Bayesian classification, artificial immune system, and rough sets, and discussed how they could be used to solve the classification problem for spam e-mails. Also, the author compares the algorithms' results on the Spam Assassin spam corpus and provides descriptions of each algorithm.

[2] The automatic categorization of papers has become essential, according to authors, especially with the rapidly expanding amount of documents available online. In order to automatically associate a document with a category, automatic categorization assigns a category as a text depending on their data it contains. The author went on to say that automatic categorization can resolve a number of issues, including determining the language of the document, screening, identifying spams, and routing and forwarding emails to their recipients. The findings of categorising Arabic text using Artificial neural networks, support vector machines, and a combined approach called BSOCHI SVM were then reported by the author. Two forms of representation root based stemming and light stemming, were used by author to explain the methodology, show the results of implementation, and summarise evaluation findings. Several performance measures were used in each example to evaluate Open Source Arabic Corpora.

[3] A learning classifier's performance may suffer from having too many features, according to writers the computing time required for training may be prolonged. In order to accelerate computation and improve classification accuracy, the author emphasised the importance of a pre-processing stages that involves feature extraction and feature reduction techniques. The author talked about the issue that was taken into consideration for the study, which has to do with data transformations prior to machine learning classifier. For recognising spam, the author suggested a feature representation with a lower dimensional space that protects class separability. According to the author, the main benefit of the suggested Feature representation is its robustness, which enables classifiers like random forest, support vector machines, and the decision tree to categorise the incoming emails as spam or ham with a small feature size and good generalisation, regardless of the data source.

[4] Spam that occurs in emails is the most well-known type, but there are various types of spam that can be found in other media. According to the author, spam 2.0 is the spread of unwanted, anonymous mass content intended to harm reputable Web 2.0 applications. Spam 2.0 can take the form of a phoney, attractive social networking profile, a promotional review, a response to an unwanted post in an online forum, or a manipulated Wiki article, for example. The authors of the paper provide an in-depth analysis of contemporary methods for filtering Spam 2.0, including detection-based, prevention-based, and early detection-based techniques.

4. Model used

Navie Bayes

Naive bayes is a machine learning algorithm that is straightforward, efficient, and frequently used. It is a probabilistic classifier that uses the maximum. A Posterior judgment algorithm in bayesian framework to classify data.

$$\mathbf{P(X/Y) = P(Y/X) * P(X) / P(Y)}$$

Eqn 1. Naïve Bayes equation

$P(Y)$ = Prior probability of Y

$P(X)$ = Prior probability of class X

$P(X|Y)$ = Occurrence of predictor X given class Y probability

Multinomial Navie Bayes

The Multinomial Naive Bayes algorithm is the most often used statistical learning method in Natural Language Processing. (NLP). The Bayes theorem serves as the foundation for the algorithm, which predicts the tag of a text such as an email or news article. It calculates the probabilities of each tag for a specific sample and outputs the tag with the highest probability.

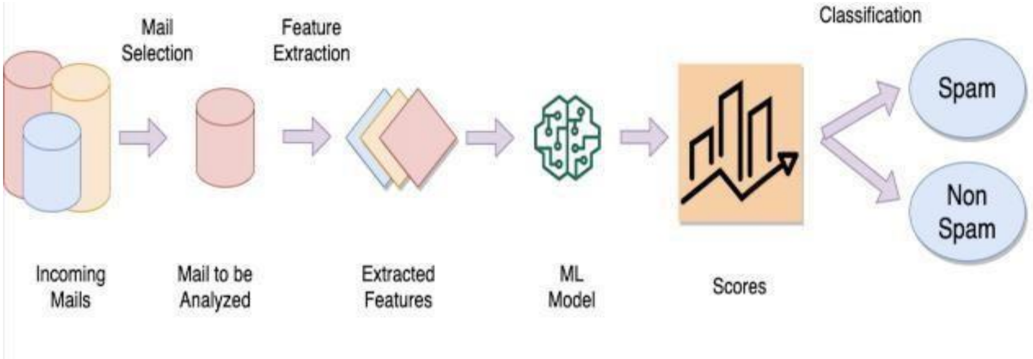

Fig 2. Architecture diagram

Workflow: Fig 3.1

1. Data Pre-processing: The first step is to pre-process the data, which involves cleaning and transforming every raw data into a format that can be used by the algorithm. This step may involve removing stop words, stemming, and converting text data into numerical data.
2. Splitting Data: Once the data is pre-processed, it split into training and testing datasets. The training dataset is used to train those classifiers, and the testing dataset is used to evaluate their performance.
3. Feature Extraction: Next, the important features are extracted from the preprocessed data. In spam detection, these features could include the presence of certain keywords, the length of the email, the use of capital letters, and so on.
4. Training: The training process involves calculating the probabilities of each feature given the spam or non-spam class. These probabilities are calculated using Bayes' theorem, which assumes that each feature is independent of the others. The probability of an email being spam is then calculated based on the probabilities of its features.
5. Classification: Once the classifier is trained, used to classify new, unseen emails as either ham or spam. This is done by calculating the probability of each email belonging to the spam or non-spam class, and choosing the class with the higher probability.
6. Evaluation: Finally, the performance of the classifiers is evaluated on the testing dataset. This involves calculating metrics such as accuracy, precision, recall, and Fl - score, which give an idea of how well the classifier is able to distinguish between spam and non-spam emails.

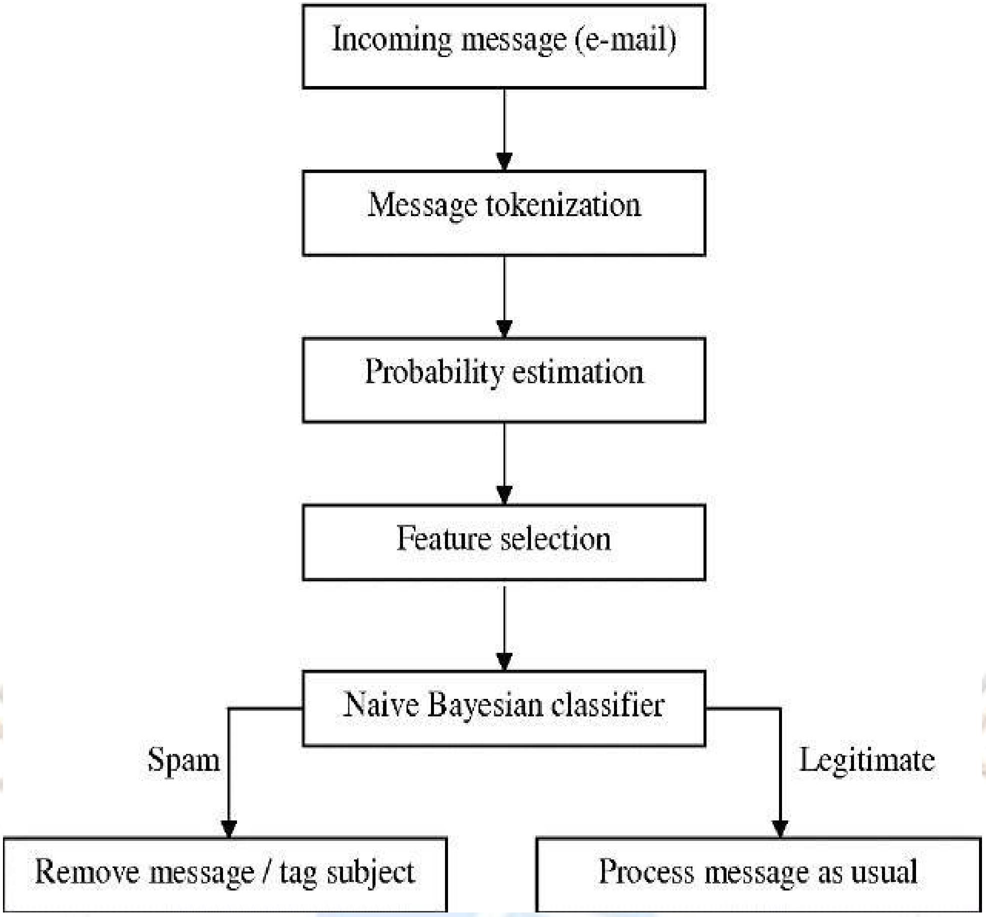

Fig 3. Workflow

5. Result

DATA SET**:**

Training set size: 4179

Testing set size: 1393

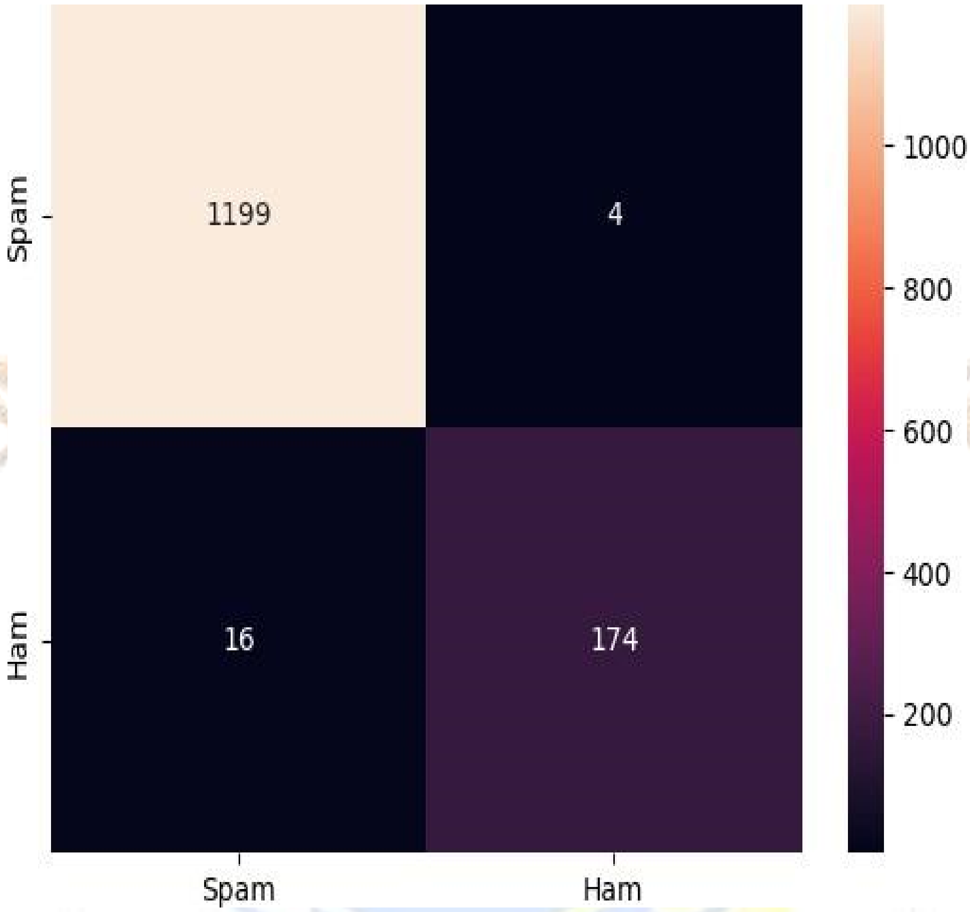

Fig 4. Confusion Matrix

6. Conclusion

The machine learning method (Multinomial Nave Bayes) is used to identify spam emails. Today, email is the most significant form of contact because it allows for global message delivery thanks to internet access. With the help of this method, junk mail will be less frequent and detected more accurately. The grid of uncertainty Image 4. Describe how the categorization algorithm performed using the test data.

In the future, this system can be applied using various methods, and it can also get more features added to it.

References

[1] W.A. Awad and S.M. ELseuofi(2011). Machine learning methods for spam email classification. Computer science and Information technology. 10.5121/ijcsit.2011.3112

[2] Riadh Belkebir and Ahmed Guessoum A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization(2013) . Computer science and application. 10.1109/AICCSA.2013.6616437

[3] Diale, M., Celik, T., & Walt, C. V. D. (2019). Unsupervised feature learning for spam email filtering. Computers & Electrical Engineering, 74, 89–104. 10.1016/j.compeleceng.2019.01.004.

[4] Vidyasagar Potdar and Pedram Hayati**.** Spam 2.0 state of the art. Digital science and Forensics. 10.4018/jdcf.2012010102

[5] Feng, Y., Chen, H., Li, T., & Luo, C. (2020). A novel community detection method based on whale optimization algorithm with evolutionary population. Applied Intelligence, 1–20. 10.1007/s10489020- 01659-7.

[6] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., &Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. IEEE Access, 7, 168261– 168295. 10.1109/ACCESS.2019.2954791.

[7] Luo, J., & Liu, Z. (2019). Novel grey wolf optimization based on modified differential evolution for numerical function optimization. Applied Intelligence, 1–19. 10.1007/s10489-019- 01521-5.

[8] Pan, X., Xue, L., & Li, R. (2019). A new and efficient firefly algorithm for numerical optimization problems. Neural Computing and Applications, 1445–1453. 10.1007/s00521-018-3449-6.

[9] Sekh, A. A., Dogra, D. P., Kar, S., Roy, P. P., & Prasad, D. K. (2020). ELM-HTM guided bio-inspired unsupervised learning for anomalous trajectory classification. Cognitive Systems Research, 63, 30–41. 10.1016/j.cogsys.2020.04.003.