International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Recent Advances in Computer Applications and Information Technology
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Spam Email Detection Using Machine Learning and Natural Language Processing Techniques

Harsh Dhakate, Mithul Jamgade

G H Raisoni University, Amravati, Maharashtra, India

Abstract

Email is one of the most crucial forms of communication in daily life, be it academic or business. Yet with the growing volume of email, comes a growing amount of spam emails. Spam emails are not only annoying but they are also dangerous because it may have links for phishing and other malicious activities as well. The strategy for detection using filter such as spam email detection does not work the same way because perpetrators change their strategies to avoid filters. So, its need of the hour to develop intelligent and automated spam email detection systems which learn and adapt.

In this research, we design and build a model for Spam Email Detection using the Machine Learning and Natural Language Processing techniques. This research aims to develop a model that can classify emails based on email content using spam and ham emails. Natural Language Processing is also used because email data is pre-processed. Pre-processing — Various methods used to clean and normalize raw email data These include tokenization, stop word removal, lower casing and lemmatization.

After pre-processing, feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) are used to convert the raw email data into a representation that machine learning algorithms can comprehend and work with. Various supervised machine learning models including Naïve Bayes, SVM, Logistic Regression and Random forest are implemented followed by performance comparison.

The models are trained on labeled datasets and then the performance of the models is evaluated using various metrics like accuracy, precision, recall, F1 score. Through this experiment we can figure out that the classification of spam messages or accuracy is highly improvised in machine learning models w.r.t. old strategies. For instance, SVM and Logistic Regression models perform well with high precision and low false positive rates so that normal messages are not predicted to be spam.

The findings of this study validate that with the help of defining higher-level features using NLP models, we can achieve improved accuracy by combining machine learning models to form a robust spam filter model, which could be scaled upwards for adapting to the changing characteristics of spam messages as well as providing useful performances in real-time. Future Directions The next step can combine deep learning models with powerful word embeddings that increase the performance of the model in spam message classification. In this paper we demonstrate the importance of intelligent methods for safe email communication models.

KEYWORDS*: Spam Email detection , Machine Learning, Natural Language Pro Naïve Bayes, TF-IDF , Support Vector Machine, Text Classification , Cybersecurity, Email Filtering, Data Mining, Email Security, Data Preprocessing, Text Vectorization, Pattern Recognition, Modelling, Intelligent Spam Filtering.*

1. Introduction

Electronic mail, also called email, is one of the most popular forms of communication in the modern world. Using email, anyone in the world can instantly send and receive messages. This has made email the fastest, most efficient, and cost-effective means of communication in the world [1]. This is the reason why email is the most popular means of communication among individuals, organizations, and businesses. This is also the reason why email is used for various purposes, including official communication, academic discussions, marketing, and even personal communications [2]. The growth of the internet and the development of email have increased the number of email users significantly over the past decade. Today, billions of emails are sent and received every day, making email the most important feature of the internet.

Despite the many advantages offered by email, the popularity of email has also resulted in many problems, the first being the emergence of the spams sent through email. Spam is the unwanted email sent to the email accounts of the masses without their consent [3]. This email is often sent for the purpose of advertisements, marketing, and even fraud. This is the reason why the email sent is often misleading and contains information that is harmful to the user or the recipient of the email.

Spam emails, apart from causing inconvenience to the users, also pose major threats to the security of the computer systems. Most spam emails have the potential to trick the users into revealing their passwords, bank account information, or other identification details. [1] Such types of attacks, where the users are tricked into revealing their information, are commonly known as phishing. Moreover, spam emails also have the potential to carry malware, which is harmful software that can attack the computer systems if the users click on the spam emails. Such types of attacks have the potential to cause major threats to the computer systems, resulting in identity theft, financial fraud, etc. Therefore, the need to protect the email systems from spam emails is now an important requirement for ensuring the safety of the computer systems [2].

Traditionally, the email service providers have employed basic types of spam filtering systems. Such types of spam filtering systems, which were employed in the early days, included rule-based systems, keyword filtering, etc. For instance, if the spam email contained words such as “free,” “win money,” or “limited offer,” the spam filtering system would mark the email as spam. [3] Such types of spam

filtering systems, which were employed in the early days, have now become outdated.

One such technique is using binary classification, where each email message is classified into two categories, i.e., spam emails or legitimate emails, also known as "ham" emails. Machine Learning (ML) techniques are quite powerful in solving classification problems. Stemming is another technique used in NLP, where a word is converted into its root word, so all the variants of a word are treated as a single feature [4]. After applying these techniques, the preprocessed data must be converted into a numerical representation so that machine learning algorithms can process this data. Techniques such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are used to extract features from the preprocessed text data. TF-IDF assigns a weight to each word, depending upon the frequency of the word in a particular email [5].

After the features are extracted, various machine learning algorithms can be used to classify the email. Various supervised learning algorithms have been used to detect spam email in the past. These algorithms are Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, etc. Each of these algorithms has its own benefits Random Forest is an ensemble learning algorithm that is used to handle high-dimensional features [6].

In this research, the focus is to develop a system to detect spam email using Machine Learning and Natural Language Processing techniques. The objective of this research is to develop different machine learning models to compare the performance of the algorithm that performs the best in classifying the spam email [7]. The dataset used in this research contains labeled email data that is divided into two classes: spam or ham. The textual features are processed using Natural Language Processing techniques, then converted into numerical features using the TF-IDF technique.

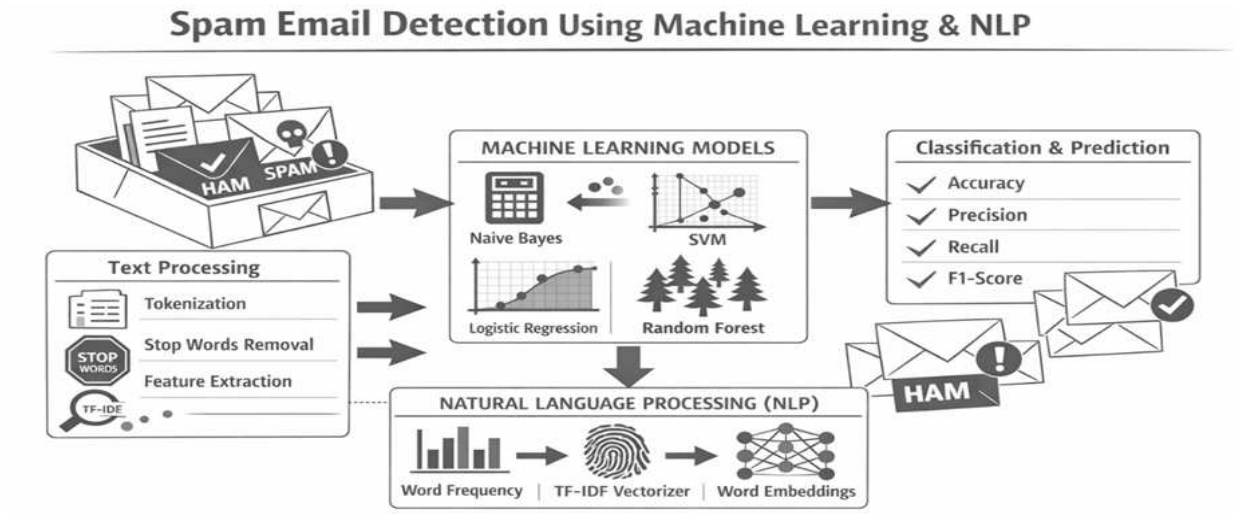


Fig.1 Spam Email detection using Machine Learning & NLP Framework

2. Literature Review

The detection of spam emails is a significant research issue that has remained relevant for a long period of time because of the potential security risk and productivity loss it poses. The traditional spam email detection systems were mainly based on rule-based systems. These systems involved checking for specific keywords such as "free," "offer," "win money," and so on [6]. These systems could also be implemented by blocking emails from suspicious senders. These systems were simple and could be implemented easily; however, they were not efficient in the long term. This is because spammers learned how to evade such systems by making minute changes such as spelling variations and additions of special characters in the email body.

To improve upon these systems and make them efficient in spam email detection, researchers introduced statistical approaches. The most widely used statistical approach was Naïve Bayes. Research studies revealed that Naïve Bayes was efficient compared to keyword-based systems because it calculated probabilities of email spam based on word frequencies. However, it was found to be inefficient in some complex situations and context-based spam emails [7].

Recently, advanced approaches such as Support Vector Machines (SVM) were introduced and found to be efficient in spam email detection systems because of their efficiency in handling high-dimensional text data. Research studies revealed that SVM is highly efficient compared to Naïve Bayes because of its higher accuracy in email spam detection systems using techniques such as TF-IDF for feature extraction [8]. The second approach, which is also widely used, is the Logistic Regression model due to its simplicity and high accuracy in email spam detection systems .The researchers have also tried the ensemble methods, namely Random Forest and Gradient Boosting, where the combination of decision trees is used to make their models more accurate .These models were less prone to overfitting and had high accuracy on various datasets; however, they were consuming more computational power compared to the other models.

With the advent of Natural Language Processing (NLP) techniques, the researchers have also started working on more accurate preprocessing techniques [4]. The researchers have observed that techniques like tokenization, stop words removal, stemming, and lemmatization can significantly increase the accuracy of their models by preprocessing the email data.

The feature engineering techniques, namely Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), are the most common techniques used to convert the email data into a numerical format that can be used in

the model [5]. TF-IDF has shown promising results in this area, where more weight is given to the unique words that are more important in the context of email spam filtering.

In recent years, researchers have started using deep learning models as well to solve the spam filtering problem. In this field, the models such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have been successful, as they can easily learn complex patterns and relationships [6]. In recent times, models named as BERT have been successful in this field, as they can easily understand the context in which words are being used, rather than looking at the words individually.

It can be clearly seen that researchers have moved from using simple rule-based models to using more sophisticated machine learning and deep learning models to solve this problem. Most researchers have shown that using more accurate preprocessing techniques with more accurate machine learning models can significantly increase the accuracy of their models in the context of spam filtering. Though most researchers have shown promising results in this area with high accuracy rates, there are still many challenges that remain to be solved in this area [7].

3. Research Methodology

In this study, we created a spam email detector using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The objective is to create a computer program that can learn to automatically determine if an email is spam or not spam, i.e., legitimate or 'ham.' The whole process in the spam email detector has been conducted step by step [9]. We first have a list of labeled emails, which means that all emails included in this list are considered to be either spam or not spam. This is important to the computer program will learn from these emails. The list of labeled emails is divided into two parts: 80% is used to train the computer program, and 20% is used to test it.

3.1. Dataset Collection

The first step in carrying out this research is to collect the data, the data collected include emails. The data collected for the research include a huge amount of emails, and this data is already classified into two types: spam emails and non-spam emails.Each email is labeled as the belonging to the category of spam emails or the category of non-spam emails [10]. This data is important for carrying out research based on machine learning , as it is labeled, i.e., it is classified into two types: spam emails and non-spam emails.In order to test the models, this data is classified into two types. 80% of the email data is used to train the models, and the remaining 20% is used to test the models.

3.2. Text Preprocessing

Email messages also contain various kinds of information that are not needed, such as punctuation marks, numbers, special characters, and common words that do not contribute much to the model for spam email detection. Using this raw data will also affect the model's performance [1]. Thus, the email data was subjected to a preprocessing step to clean the data before it is fed into the model.

The first step in the preprocessing step is converting the email data into lower case. This is to ensure that words such as “Free” and “free” are treated as the same word, ensuring the data is consistent throughout the model. The second step is removing punctuation marks, numbers, and special characters, as this information is usually not useful for the model to classify the email as spam or not [2].

The second step is removing stop words from the email data. Stop words are common words that appear in almost all sentences, such as “the,” “is,” “and,” “in,” “of,” and many more. This step is important in the preprocessing step because the model will not benefit from the presence of these common words when classifying the email as spam or not.The third step is lemmatization, whereby the word is converted to its base or root word. For instance, the words “running,” “runs,” and “ran” will be converted to the base word “run ” [3].

3.3. Feature Extraction

The second step in the model is converting the email data into numerical features, as the model will not understand the email data in the original format.In this research, the Term Frequency-Inverse Document Frequency technique is employed for feature extraction. This is a commonly used technique for feature extraction in text classification tasks, as it assists in determining the significance of words in a document relative to the entire dataset [2].

Term Frequency is used to determine how frequently a word is used in a particular email, whereas Inverse Document Frequency is used to determine how important a word is in the dataset. This technique will assign higher weights to words that frequently appear in an email but do not frequently appear in other emails [3].

This technique will ensure that words with higher occurrence in an email, but with lower occurrence in other emails, are assigned higher weights.This technique will ensure that emails are converted into vectors, making it easier for machine learning techniques to classify the emails, thereby highlighting the words with higher chances of being used by spammers [4].

3.4. Model Training

The extracted features will then be used to train machine learning classification algorithms on the dataset. In this research, various machine learning classification algorithms have been employed to train the dataset. Four different models have been created to compare their performance [5].

The classification algorithms used in this research are as follows:

Naive Bayes: This is a classification algorithm that uses probability to determine how likely it is for an email to belong to a particular category based on words used.

Support Vector Machine (SVM): This is a Good machine learning algorithm that creates an boundary between emails that belong to the spam category and emails belonging to other categories.

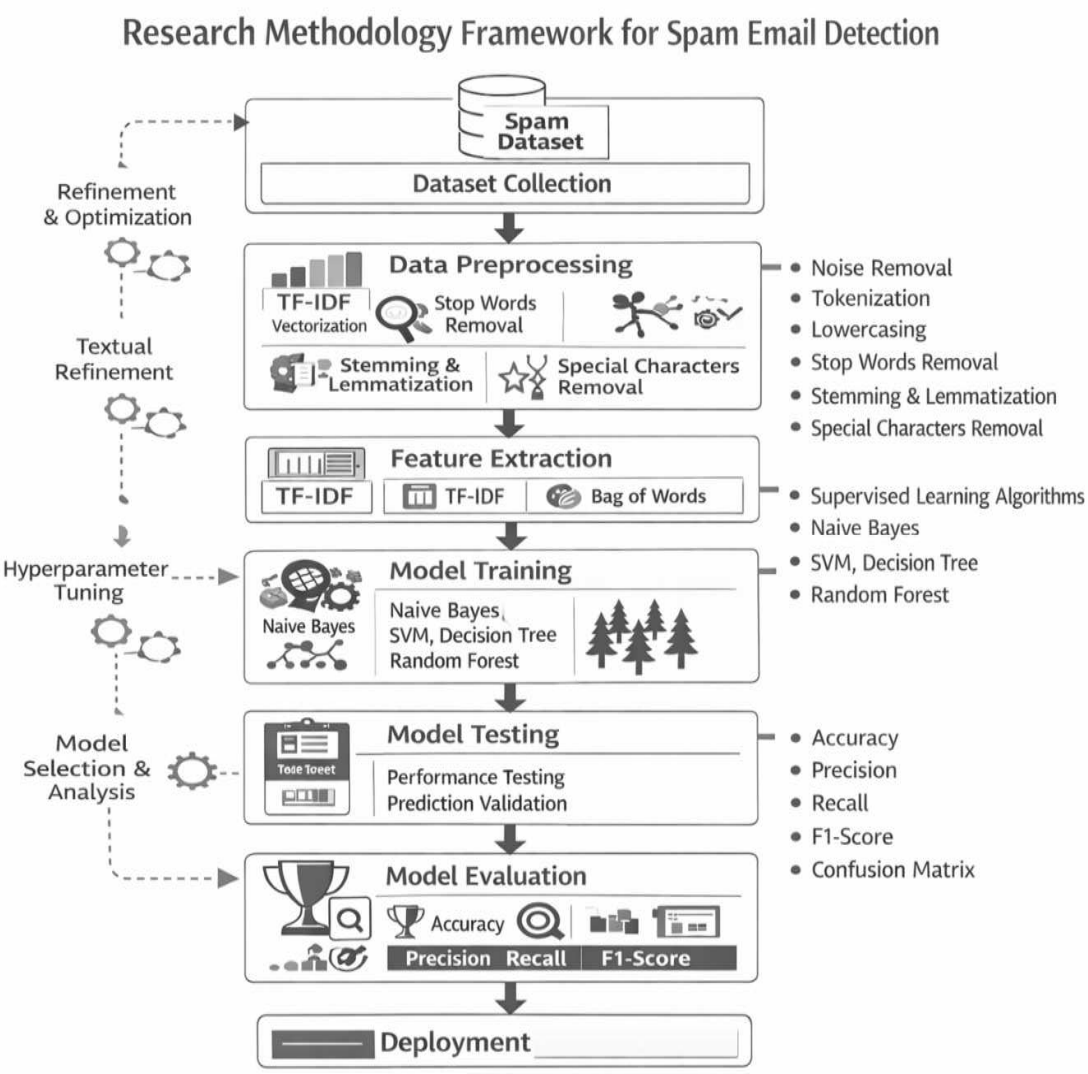


Fig 2. Proposed research methodology framework

Logistic Regression: This is a statistical technique used for binary classification, where the likelihood of an email being either spam or non-spam will be predicted [6].

Random Forest: This is a technique used for machine learning, where emails will be classified using a combination of decision trees.

3.5. Model Evaluation

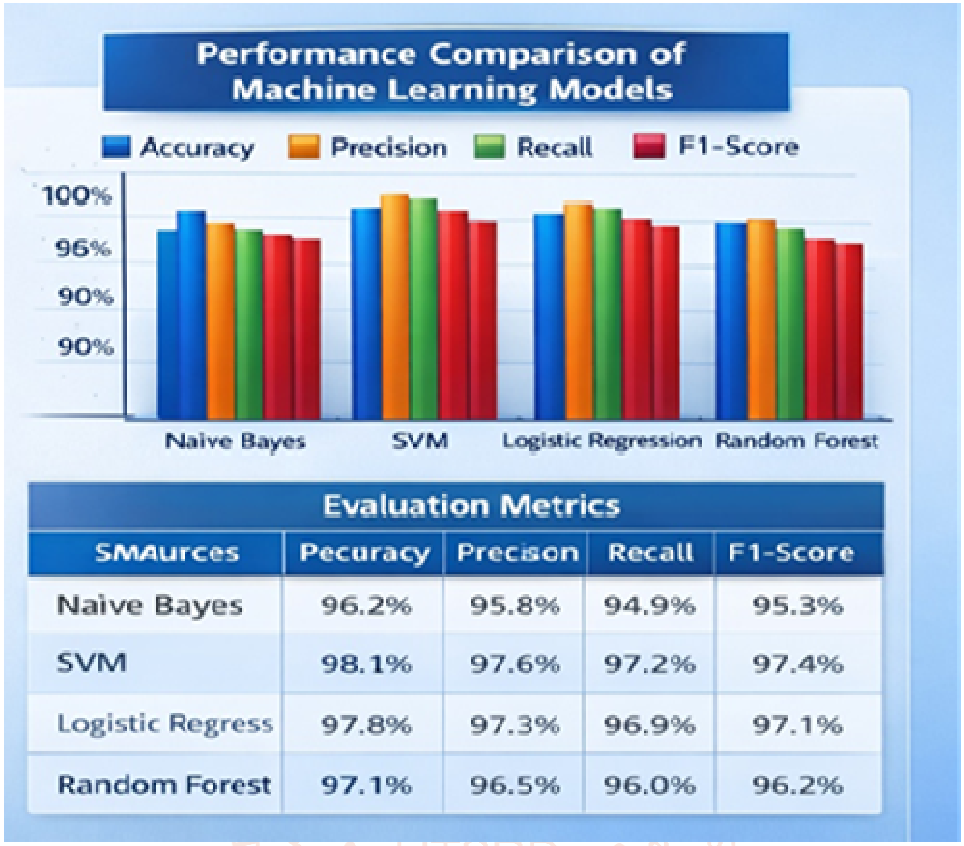
After training the machine learning models, their performance is then evaluated using the testing dataset. This step is used to evaluate the performance of the trained machine learning models to establish the level of accuracy with which the models can classify new emails that they have not previously seen [10].The evaluation metrics used to evaluate the performance of the machine learning models are : Accuracy, Precision, Recall, F1-score.

Accuracy is used to evaluate the level of correctly classified email messages by the machine learning models. Precision is used to evaluate the number of email messages that are classified as spam. Recall is used to evaluate the number of email messages that are classified as spam by the machine learning models. The F1-score is used to evaluate the performance of the machine learning models in a more comprehensive manner [11].

4. Result

The performance evaluation of the proposed spam email detection system was carried out using several machine learning algorithms. In this process, the dataset was split into training sets and testing sets, where 80% of the dataset model was used to train the models, and the remaining 20% of the dataset was used to monitor the performance of the models. the training of the models, the email text data was subjected to Natural Language Processing techniques such as tokenization, removal of stop words, lowercasing, and lemmatization [13]. After the preprocessing of the email text data, the features of the email data were extracted using the TF-IDF method.

In this study, four machine learning algorithms are proposed to detect spam email. With the help of this four machine learning algorithms used in this study are Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest. Each of the machine learning models was trained by using the email dataset, and the performance of the models was evaluated using several performance metrics such as the Accuracy, Precision, Recall, and F1-Score [12]. The evaluation metrics will help to better understand the performance of the machine learning models in classifying the spam email correctly.



Evaluation Metrics

SMAurces	Pecuracy	Precison	Recall	F1-Score
Naive Bayes	96.2%	95.8%	94.9%	95.3%
SVM	98.1%	97.6%	97.2%	97.4%
Logistic Regress	97.8%	97.3%	96.9%	97.1%
Random Forest	97.1%	96.5%	96.0%	96.2%

Fig 3. Output of Spam Email detection system

From the results obtained, it is a proof that all the machine learning models performed well in the classification of the spam emails, though the performance of the models differs based on the approach used in the model. The Naïve Bayes model was performing well in the classification of the spam email. This is due to the fact that the model is efficient in the classification of the text-based data [14]. The model was able to classify the spam emails with relative accuracy in a short time.

The Support Vector Machine model was performing well in the classification of the spam emails compared to the Naïve Bayes model. This is due to the fact that the model is effective in the high-dimensional space, and it is efficient in the classification of the text-based data. The model was able to differentiate between the spam and non-spam emails [9].

5. Conclusion

Email is really important to us in our lives. We use email for things for school and for work. With all the other ways we can talk to each other online email is still the best way to communicate.. There is a big problem with email. Lots of people get emails that they do not want these are called spam emails. Spam emails are an issue because they can hurt our computers and steal our personal information [15].

Email spam can lead to things, like phishing which's when someone tries to trick us into giving them our secrets. It can also give our computers software, which is called malware. Sometimes spam emails can even steal our identity. Take our money, which is very bad. Conventional rule-based email filtering systems have proved to be ineffective in dealing with the growing spams [16]. In recent times, spammers have adopted different ways to trick email filtering systems. For instance, they make use of misleading phrases, special characters, as well as content modifications to trick email filtering systems.

In view of this, there is a dire need to adopt intelligent email filtering systems. In this research work, it has been proved that Machine Learning (ML) coupled with Natural Language Processing (NLP) can be used as an effective as well as reliable solution for filtering out spam emails. In this research work, it has been proved that the binary classification problem of spam detection can be solved with the help of Machine Learning. In this regard, NLP can be used to clean up the email messages [17]. Feature extraction techniques like TF-IDF play a major role in identifying the key words in email messages.

- It reduces the number of spam emails that end up in your inbox.
- It minimizes the number of positives, which are legitimate emails that get marked as spam.
- The Support Vector Machine model and the Logistic Regression model are really good at filtering out spam emails.
- The machine learning system is effective, in keeping your inbox clean.

REFERENCES

[1] T. Almeida, J. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in *Proceedings of the 11th ACM Symposium on Document Engineering*, Mountain View, CA, USA, 2011, pp. 259–262.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.

[3] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in

Proceedings of the European Conference on Machine Learning, Berlin, Germany, 1998, pp. 137–142.

[4] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, Madison, WI, USA, 1998, pp. 41–48.

[5] G. V. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 335–455, 2008.

[6] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proceedings of the 23rd Annual International ACM SIGIR Conference*, Athens, Greece, 2000, pp. 160–167.

[7] P. Graham, "A plan for spam," *MIT Technology Review*, vol. 107, no. 2, pp. 1–10, 2004.

[8] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of Naive Bayes text classifiers," in *Proceedings of the 20th International Conference on Machine Learning*, Washington, DC, USA, 2003, pp. 616–623.

[9] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002.

[10] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems*, vol. 36, pp. 226–235, Dec. 2012.

[11] S. S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*. Boston, MA, USA: Springer, 2016.

[12] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, USA, 1997, pp. 412–420.

[13] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[14] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2–3, pp. 103–134, 2000.

[15] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*, Piscataway, NJ, USA, 2003, pp. 133–142.