IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Spam Email Detection Using TF-IDF Based Support Vector Machine Approach

Dr. K. Kasturi[1] and Dr. Y. Kalpana[2]
[1]Associate Professor & [2]Professor, Department of Applied Computing and Emerging Technologies,
School of Computing Sciences, VISTAS, India
kasturi.scs@vistas.ac.in[1], kalpana.scs@vistas.ac.in[2]

Abstract: *Nowadays, all the people are communicating official information through emails. Spam mail is the major issue on the internet. It is easy to send an email which contains spam messages by the spammers. Spam fills our inbox with several irrelevant emails. Spammers can steal sensitive information from our device like files, contact. Even though we have the latest technology, it is challenging to detect spam emails. This paper aims to propose a Term Frequency Inverse Document Frequency (TFIDF) approach by implementing the Support Vector Machine algorithm. The results are compared in terms of the confusion matrix and accuracy. and precision. This approach gives an accuracy of 99.9% on training data and 98.2% on testing data achieved by using the Term Frequency Inverse Document Frequency (TFIDF) based Support Vector Machine (SVM) system.*

Keywords: Ham/spam, Natural Language Processing, Machine Learning, Online Platform, Email, TFIDF.

I. INTRODUCTION

Email spam has become a significant problem in today's digital age, posing challenges for individuals, businesses, and organizations alike. Spam emails are unsolicited messages that flood inboxes, wasting valuable time and resources while potentially exposing users to malicious content or scams. To combat this issue, machine learning techniques have emerged as powerful tools for email spam detection.

There are several machine learning techniques commonly employed for email spam detection. These include Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and Neural Networks. These algorithms can be trained on labeled datasets, allowing them to learn the underlying patterns and relationships between spam and non-spam emails. The success of email spam detection using machine learning heavily relies on the quality and diversity of the training data. A comprehensive dataset that covers a wide range of spam types and legitimate emails is essential for training robust models.

Additionally, feature engineering plays a crucial role in Identifying relevant attributes and extracting meaningful information from email data. The benefits of using machine learning for email spam detection are numerous. It enables efficient filtering and separation of legitimate emails from spam, reducing the time and effort spent by users in manually sorting through their inbox. Moreover, machine learning models can adapt to evolving spam techniques, continuously improving their accuracy over time.

Email is one of the most popular communication methods, but unfortunately, it is also a common target for spam messages. Spam emails not only waste time but can also contain malicious links or attachments that can harm computer systems. As the volume of emails continues to grow, it has become challenging to identify and classify spam emails manually. Therefore, the development of machine learning (ML) and natural language processing (NLP) techniques has opened new avenues for automated email spam classification. In this project, we aim to use ML and NLP techniques to classify emails as spam or legitimate, based on their content and their relevant features. The project involves building a model to analyze the text of emails and determine whether they are spam or legitimate. This study has the potential to provide a valuable solution to the problem of email spam and help users to manage their emails more effectively.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

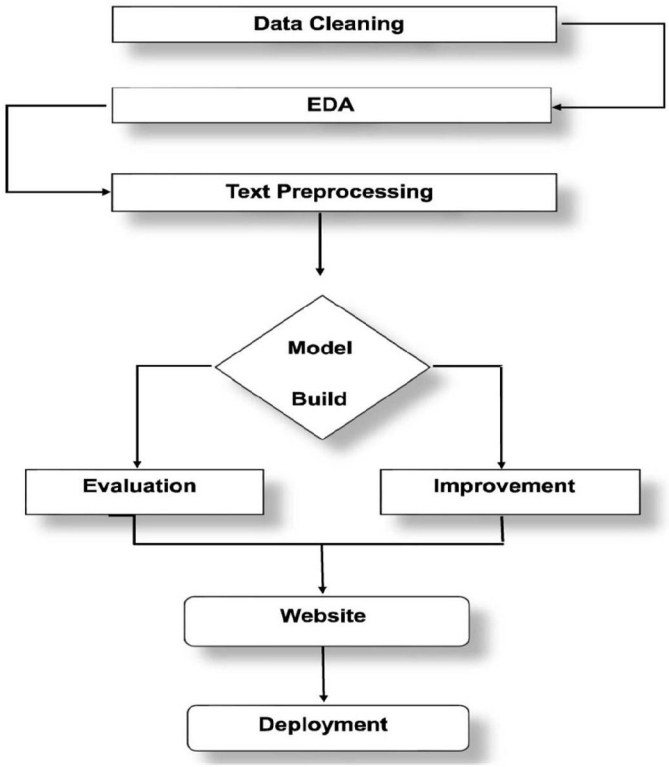


Fig 1.1 : Flow Diagram

1.1 METHODOLOGY:

The era of creation of this product includes models i.e., object-oriented model, Prototype model, waterfall model etc. for making the correct system. Water model, the oldest model of creation of correct system. The product model used by our framework is the cascade model. Cascade model could be a precise and successive way to contend with the merchandise improvement. This incorporates framework coming up with and displaying that sets up requirements for all the framework parts and distributes some set of those conditions to programming. Framework building and examination incorporate requirement gathering at the framework level with modest amount of top-ranking arrange. Examination info building consolidate would like assortment at the key business level and at the business space level.

1.2 EXISTING SYSTEM:

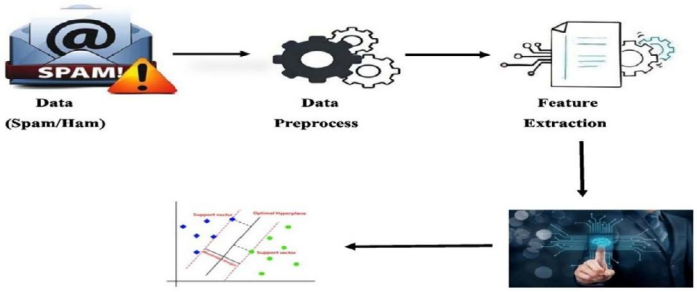
- ✓ Email Spam Classifier based on Machine Leaning Techniques had done by using SVM, KNN, Naive Bayes and Decision tree algorithms etc.
- ✓ SVM had an average accuracy of 99.6%.
- ✓ It had good accuracy when compared to the other algorithms in proposed system.

1.3 PROPOSED SYSTEM:

1. E-mail Spam Classifier is used to classify email data into spam and ham emails.
2. This method is performed by using Support Vector Machine (SVM) algorithm.
3. In this method, dataset is divided into two sets based on labels and given as input to algorithm.
4. The accuracy of 99% on training data and 98.2% on test data is obtained through the proposed system.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

1.4 ARCHITECTURE DIAGRAM:



II. DATA COLLECTION AND PREPROCESSING

Data collected from the Kaggle data set. This data set contains a collection of messages, which can be a mix of both spam and non-spam (ham) messages. The primary purpose of this data set is to develop and test algorithms or models for distinguishing between spam and non-spam messages. It is commonly used in natural language processing and machine learning tasks related to text classification.

2.1 DATA CLEANING

Data cleaning is a critical step in preparing a spam classifier data set for training machine learning models. Cleaning the data helps ensure that the model can learn effectively from the information provided. Here are some common data cleaning steps for a spam classifier data set

2.1.1 RENAMING THE COLOUMS:

In this code, replace 'old_column_name1' with the name of the column you want to change, and 'new_column_name1' with the new name you want to assign. You can repeat this for multiple columns as needed.

2.1.2 CHECK THE MISSING VALUE:

The isnull() function is used to identify missing values, and sum() then calculates the number of missing values in each column. The result will be a Series that shows the count of missing values for each column in your dataset.

2.1.3 CHECK THE DUPLICATE VALUE:

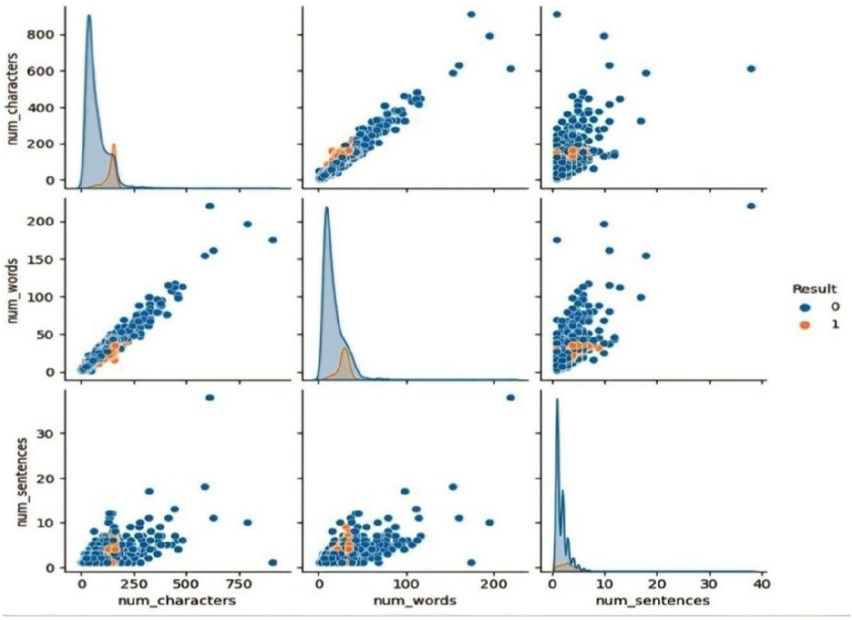
The duplicated() function identifies rows that are exact duplicates in the Data Frame. To remove duplicates from the Data Frame, we can use the drop_duplicates() function.

III. EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) for a spam Classifier typically involves understanding the characteristics of your dataset, identifying patterns and relationships in the data, and gaining insights that can inform the design of the classifier.

DOI: 10.48175/IJARSCT-34685

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2



IV. TEXT PREPROCESSING

- ✓ **Lowercasing:** Convert all text to lowercase to ensure uniformity.
- ✓ **Tokenization:** Split the text into individual words or tokens.
- ✓ **Stop Word Removal:** Remove common words (e.g., "the," "and," "is") that don't carry much information.
- ✓ **Spceial Character Removal:** Remove punctuation, symbols, and non- alphanumeric characters.
- ✓ **Stemming or Lemmatization:** Reduce words to their base or root form (e.g., "running" becomes "run").

V. MODEL BUILDING

i. At the scikit-learn library (commonly referred to as sklearn), you can follow these general steps. Scikit-learn is a popular Python library for machine learning tasks.

ii. Assuming you have your data loaded into arrays or Data Frames(X for features and y for labels), here's how you can create a train- test split and train a machine learning model.

iii. Import Necessary Libraries:
You need to import the required libraries for your machine learning project. Common libraries include scikit-learn for machine learning tools, NumPy for numerical operations, and pandas for data handling.

iv. Load and Prepare Your Data:
Load your dataset and prepare it by separating the features
(X) and the target variable (y). Ensure that your data is in a format suitable for machine learning.

v. Split Data into Training and Testing model:
Split your data into a training set and a testing set. This is typically done using the train_test_split function In this example, 20% of the data is reserved for testing, and random state ensures reproducibility.

vi. Define and Train Your Model:
Choose a machine learning model and initialize it. train the model on the training data.

vii. Make Predictions on the Test Data:
Use the trained model to make predictions on the test data.

viii. Evaluate Model Performance:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Use appropriate metrics to evaluate your model's performance. For classification tasks, you can use accuracy, precision, recall, F1- score, etc.

5.1 EVALUATING MACHINE LEARNING ALGORITHMS

	Algorithm	Accuracy	Precision
0	NB	0.977735	0.952000
1	ETC	0.977735	1.000000
2	SVC	0.970958	1.000000
3	xgb	0.968054	0.981308
4	LR	0.967086	1.000000
5	AdaBoost	0.964182	0.971429
6	BgC	0.964182	0.896000
7	RF	0.963214	1.000000
8	GBDT	0.947725	0.965909
9	DT	0.943853	0.875000
10	KN	0.913843	1.000000

Voting Classisfier : A Voting Classifier in machine learning is like taking a poll among different prediction models. Imagine you have several friends, each with their own guess about an answer. A Voting Classifier combines these guesses to make a final decision. There are two types:

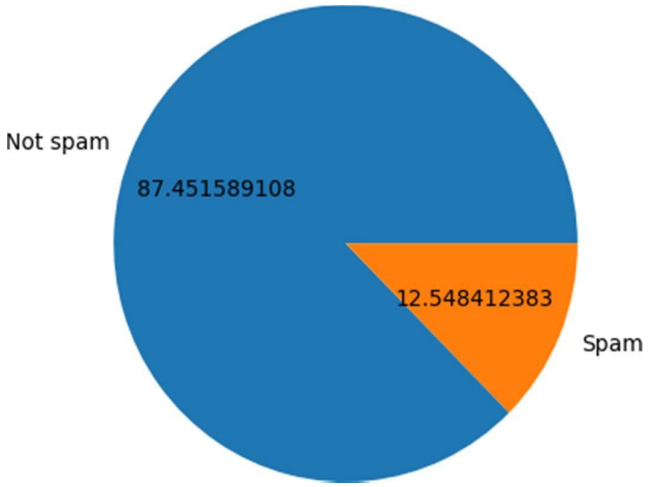
Hard Voting : It counts the majority vote to make a decision. If most friends say "yes," the final decision is "yes."

Soft Voting : It calculates the average confidence of each friend. It considers how sure each friend is and takes an average to decide.

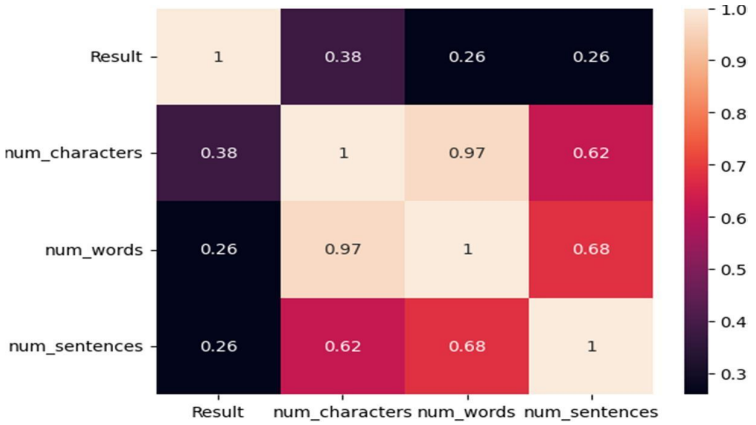
5.2 DECISION TREE CLASSIFIER RESULTS:

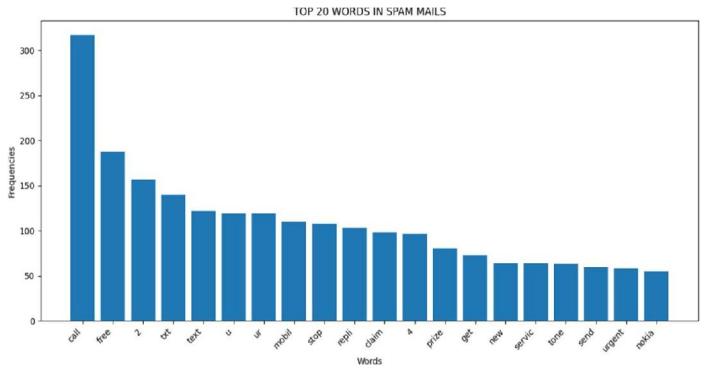
ACCURACY: 0.9816069699903195%

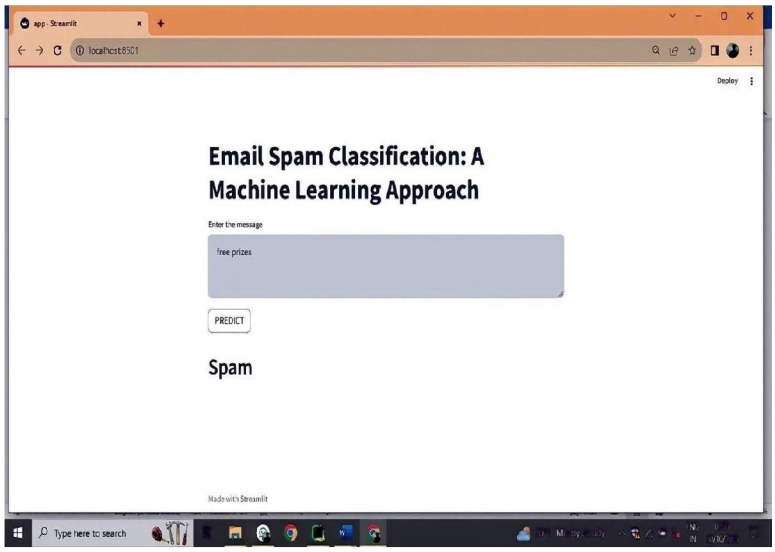
PRECISION: 0.975609756097561%



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2







5.3 CONCLUSION

Spam email is one of the most demanding and troublesome internet issues in today's world of communication and technology. It is almost impossible to think about e-mail without considering the issue of spam. Spammers by

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

generating spam mails are misusing this communication facility and thus affecting organizations and many email users. The machine learning model used by Google have now advanced to the point that it can detect and filter outs pam and phishing emails with about 99.9percent accuracy. The implication of this is that one out of a thousand messages succeed in evading their email spam filter.

5.4 FUTURE WORK

This model could be modified to work on the sender side instead of the receiver side, this way the network traffic could be reduced, and the data storage can be reduced. Also, the email IDs could have a ranking system, using this way also the above- mentioned problems could be overcome. The other methods can be that instead of the whole message being stored for analysis only the header, the attachments and the links could be analyzed. Using the above-mentioned point, the privacy of an individual could be maintained or encrypting the confidential texts that are chosen by the sender could be employed. For more accuracy the dataset of the model could be updated for the latest trends i.e., the spam and advertisements can vary on the current trends that the society is boosting at the time which will be used more to attract people by scammers.

REFERENCES

1. Ahmed Khorsi, "An Overview of Content-based Spam Filtering Techniques", Informatica, vol. 31, no. 3,October 2007, pp 269- 277.
2. Alistair McDonald, “Spam Assassin: A Practical Guide to Integration and Configuration”, 1st Edition, Packet publishers, 2004.
3. Ian H. Witten, Eibe Frank, “Data Mining – Practical Machine Learning Tools and Techniques,” 2nd Edition, Elsevier, 2005.
4. Deepika Mallampati, Nagaratna P. Hegde “A MachineLearning Based Email Spam Classification Framework Model” in IJITEE, ISSN: 2278-3075, Vol.9 Issue.4, Feb 2020.
5. Kasturi, K. "Comparison of machine learning models for diabetes prediction." International Journal of Advanced Research in Science, Communication and Technology 2 (2024).
6. J. Jamuna and K. Kasturi, "Enhancing Parkinson's Disease Prediction Using Machine Learning Techniques," 2025 9th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2025, pp. 958-964, doi: 10.1109/ICISC65841.2025.11188216.
7. Spam Assassin, “Spam and Ham Dataset", Kaggle, 2018.https://www.kaggle.com/veleon/ham-andspam-dataset.
8. Apache, “open-source Apache Spam Assassin Dataset”,2019.
9. https://spamassassin.apache.org/old/publiccorpus.
10. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.
11. Zhang, Le, Jingbo Zhu, and Tianshun Yao. "An evaluation of statistical spam filtering techniques." ACM Transactions on Asian Language Information Processing (TALIP) 3, no. 4 (2004): 243-269