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 5, Issue 8, May 2025
Impact Factor: 7.67

Spam Detection Using Machine Learning and Natural Language Processing

Dhumal Mohini, Survase Rutuja, Sonawane Pradnya
Department of Computer Engineering
Adsul's Technical Campus, Chas, Ahmednagar

Abstract: *Nowadays communication plays a major role in everything be it professional or personal. Email communication service is being used extensively because of its free use services, low-cost operations, accessibility, and popularity. Emails have one major security flaw that is anyone can send an email to anyone just by getting their unique user id. This security flaw is being exploited by some businesses and ill-motivated persons for advertising, phishing, malicious purposes, and finally fraud. This produces a kind of email category called SPAM.*

Spam refers to any email that contains an advertisement, unrelated and frequent emails. These emails are increasing day by day in numbers. Studies show that around 55 percent of all emails are some kind of spam. A lot of effort is being put into this by service providers. Spam is evolving by changing the obvious markers of detection. Moreover, the spam detection of service providers can never be aggressive with classification because it may cause potential information loss to incase of a misclassification.

To tackle this problem we present a new and efficient method to detect spam using machine learning and natural language processing. A tool that can detect and classify spam. In addition to that, it also provides information regarding the text provided in a quick view format for user convenience.

Keywords: communication

I. INTRODUCTION

Today, Spam has become a major problem in communication over internet. It has been accounted that around 55% of all emails are reported as spam and the number has been growing steadily. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chances has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world.

Spam has been a major concern given the offensive content of messages, spam is a waste of time. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation.

Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique.

II. LITERATURE REVIEW

2.1 Introduction

This chapter discusses about the literature review for machine learning classifier that being used in previous researches and projects. It is not about information gathering but it summarize the prior research that related to this project. It involves the process of searching, reading, analyzing, summarizing and evaluating the reading materials based on the project.

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 5, Issue 8, May 2025
Impact Factor: 7.67

A lot of research has been done on spam detection using machine learning. But due to the evolvement of spam and development of various technologies the proposed methods are not dependable. Natural language processing is one of the lesser known fields in machine learning and it reflects here with comparatively less work present.

2.2 Related work

Spam classification is a problem that is neither new nor simple. A lot of research has been done and several effective methods have been proposed.

i. M. RAZA, N. D. Jayasinghe, and M. M. A. Muslam have analyzed various techniques for spam classification and concluded that naïve Bayes and support vector machines have higher accuracy than the rest, around 91% consistently [1].

ii. S. Gadde, A. Lakshmanarao, and S. Satyanarayana in their paper on spam detection concluded that the LSTM system resulted in higher accuracy of 98%[2].

iii. P. Sethi, V. Bhandari, and B. Kohli concluded that machine learning algorithms perform differently depending on the presence of different attributes [3].

iv. H. Karamollaoglu, İ. A. Dogru, and M. Dorterler performed spam classification on Turkish messages and emails using both naïve Bayes classification algorithms and support vector machines and concluded that the accuracies of both models measured around 90% [4].

2.3 Summary

From various studies, we can take that for various types of data various models performs better. Naïve Bayes, random forest, SVM, logistic regression are some of the most used algorithms in spam detection and classification.

III. OBJECTIVES AND SCOPE

3.1 Problem Statement

Spammers are in continuous war with Email service providers. Email service providers implement various spam filtering methods to retain their users, and spammers are continuously changing patterns, using various embedding tricks to get through filtering. These filters can never be too aggressive because a slight misclassification may lead to important information loss for consumer. A rigid filtering method with additional reinforcements is needed to tackle this problem.

3.2 Objectives

The objectives of this project are

i. To create a ensemble algorithm for classification of spam with highest possible accuracy.

ii. To study on how to use machine learning for spam detection.

iii. To study how natural language processing techniques can be implemented in spam detection.

iv. To provide user with insights of the given text leveraging the created algorithm and NLP.

3.3 Project Scope

This project needs a coordinated scope of work.

i. Combine existing machine learning algorithms to form a better ensemble algorithm.

ii. Clean, processing and make use of the dataset for training and testing the model created.

iii. Analyze the texts and extract entities for presentation.

3.4 Limitations

This Project has certain limitations.

i. This can only predict and classify spam but not block it.

ii. Analysis can be tricky for some alphanumeric messages and it may struggle with entity detection.

iii. Since the data is reasonably large it may take a few seconds to classify and ane the message.

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 5, Issue 8, May 2025
Impact Factor: 7.67

IV. EXPERIMENTATION AND METHODS

4.1 Introduction

This chapter will explain the specific details on the methodology being used to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering. So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself.

4.2 System Architecture

The application overview has been presented below and it gives a basic structure of the application.

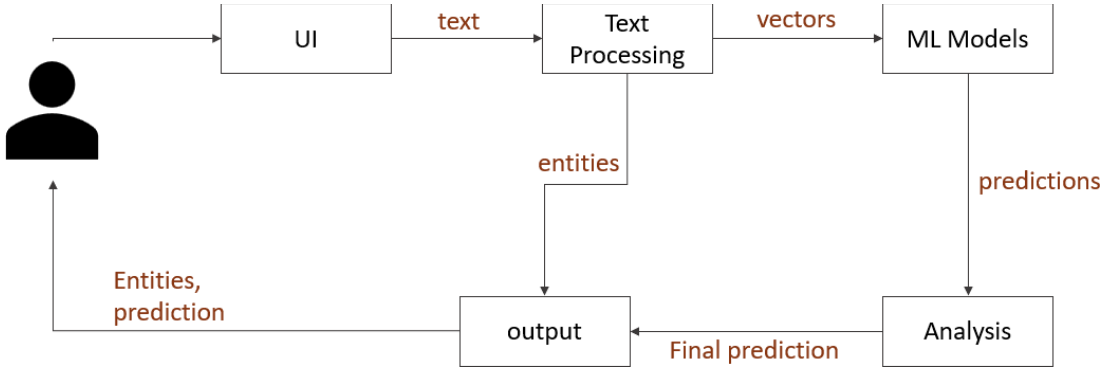


Fig no. 4.1 Architecture

The UI, Text processing and ML Models are the three important modules of this project. Each Module's explanation has been given in the later sections of this chapter.

4.3 Modules and Explanation

The Application consists of three modules.

i. UI
ii. Machine Learning
iii. Data Processing

I. UI Module

a. This Module contains all the functions related to UI(user interface).
b. The user interface of this application is designed using Streamlit library from python based packages.
c. The user inputs are acquired using the functions of this library and forwarded to data processing module for processing and conversion.
d. Finally the output from ML module is sent to this module and from this module to user in visual form.

II. Machine Learning Module

a. This module is the main module of all three modules.
b. This modules performs everything related to machine learning and results analysis.
c. Some main functions of this module are
i. Training machine learning models.
ii. Testing the model
iii. Determining the respective parameter values for each model.
iv. Key-word extraction.
v. Final output calculation
d. The output from this module is forwarded to UI for providing visual response to user

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 5, Issue 8, May 2025
Impact Factor: 7.67

III. Data Processing Module

a. The raw data undergoes several modifications in this module for further process.
b. Some of the main functions of this module includes
i. Data cleaning
ii. Data merging of datasets
iii. Text Processing using NLP
iv. Conversion of text data into numerical data(feature vectors).
v. Splitting of data.
c. All the data processing is done using Pandas and NumPy libraries.
d. Text processing and text conversion is done using NLTK and scikit-learn libraries.

4.4 Requirements

Hardware Requirements

PC/Laptop Ram – 8 Gig
Storage – 100-200 Mb

Software Requirements

OS – Windows 7 and above
Code Editor – Pycharm, VS Code, Built in IDE
Anaconda environment with packages nltk, numpy, pandas, sklearn, tkinter, nltk data. Supported browser such as chrome, firefox, opera etc..

4.5 WorkFlow

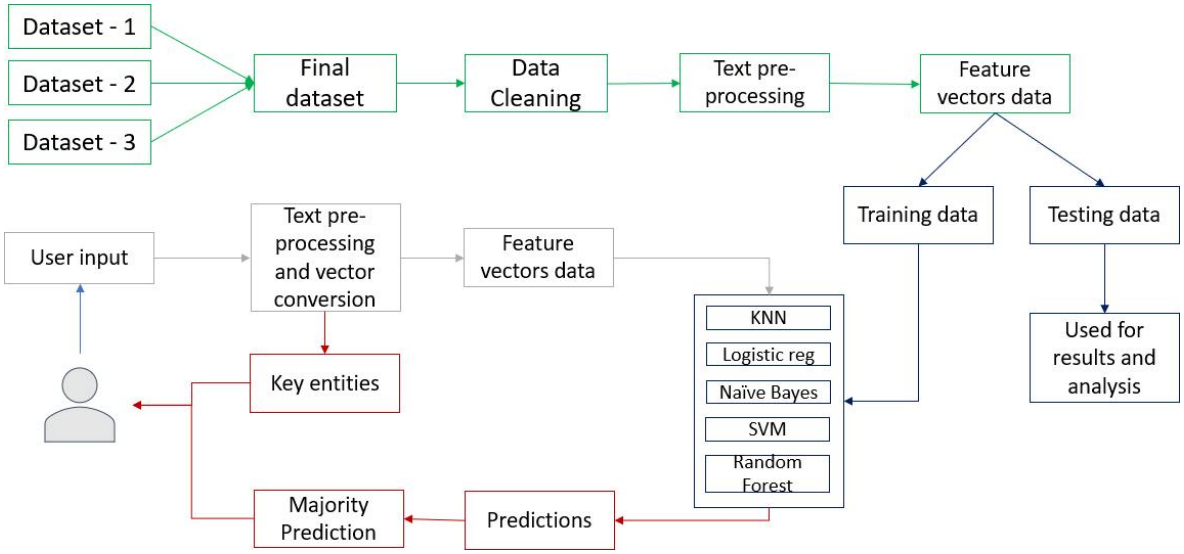


Fig no. 4.2 Workflow

In the above architecture, the objects depicted in Green belong to a module called Data Processing. It includes several functions related to data processing, natural Language Processing. The objects depicted in Blue belong to the Machine Learning module. It is where everything related to ML is embedded. The red objects represent final results and outputs.

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 5, Issue 8, May 2025
Impact Factor: 7.67

4.5.1 Data Collection and Description

- Data plays an important role when it comes to prediction and classification, the more the data the more the accuracy will be.
- The data used in this project is completely open-source and has been taken from various resources like Kaggle and UCI

Data Description

Dataset : enronSpamSubset.
Source : Kaggle
Description : this dataset is part of a larger dataset called enron. This dataset contains a set of spam and non-spam emails with 0 for non spam and 1 for spam in label attribute.
Composition :
Unique values : 9687 Spam values : 5000 Non-spam values : 4687

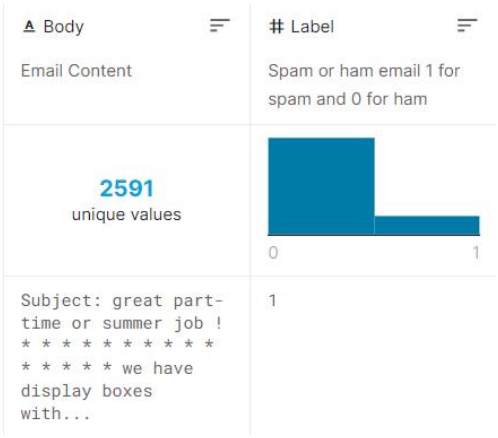


Fig no. 4.3 enron spam

Dataset : lingspam. Source : Kaggle
Description : This dataset is part of a larger dataset called Enron1 which contains emails classified as spam or ham(not-spam).
Composition :
Unique values : 2591 Spam values : 419
Non-spam values : 2172

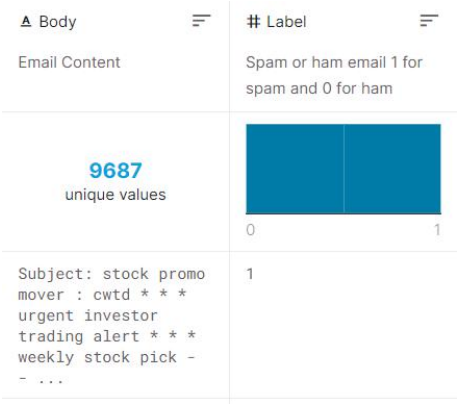


Fig no. 4.4 lingspam

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 5, Issue 8, May 2025
Impact Factor: 7.67

4.5.2 Data Processing

4.5.2.1 Overall data processing

It consists of two main tasks

Dataset cleaning

It includes tasks such as removal of outliers, null value removal, removal of unwanted features from data.

Dataset Merging

After data cleaning, the datasets are merged to form a single dataset containing only two features(text, label).

4.5.2.2 Textual data processing

Tag removal

Removing all kinds of tags and unknown characters from text using regular expressions through Regex library.

Sentencing, tokenization

Breaking down the text(email/SMS) into sentences and then into tokens(words).

Stop word removal

Stop words such as of , a ,be , … are removed using stopwords NLTK library of python.

Lemmatization

Words are converted into their base forms using lemmatization and pos-tagging

Sentence formation

The lemmatized tokens are combined to form a sentence.

This sentence is essentially a sentence converted into its base form and removing stop words.

4.5.3 Data Splitting

The data splitting is done to create two kinds of data Training data and testing data.

Training data is used to train the machine learning models and testing data is used to test the models and analyze results. 80% of total data is selected as testing data and remaining data is testing data.

4.5.4 Machine Learning

4.5.4.1 Introduction

Machine Learning is process in which the computer performs certain tasks without giving instructions. In this case the models takes the training data and train on them.

Then depending on the trained data any new unknown data will be processed based on the ruled derived from the trained data.

After completing the count vectorization and TF-IDF stages in the workflow the data is converted into vector form (numerical form) which is used for training and testing models.

4.5.4.2 Algorithms

a combination of 5 algorithms are used for the classifications.

4.5.4.2.1 Naïve Bayes Classifier

A naïve Bayes classifier is a supervised probabilistic machine learning model that is used for classification tasks. The main principle behind this model is the Bayes theorem.

Bayes Theorem:

Naive Bayes is a classification technique that is based on Bayes' Theorem with an assumption that all the features that predict the target value are independent of each other. It calculates the probability of each class and then picks the one with the highest probability.

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 5, Issue 8, May 2025
Impact Factor: 7.67

Naive Bayes classifier assumes that the features we use to predict the target are independent and do not affect each other. Though the independence assumption is never correct in real-world data, but often works well in practice. so that it is called "Naive" [14].

$$P(A \mid B)=(P(B \mid A)P(A))/P(B)$$

P(A|B) is the probability of hypothesis A given the data B. This is called the posterior probability.
P(B|A) is the probability of data B given that hypothesis A was true.
P(A) is the probability of hypothesis A being true (regardless of the data). This is called the prior probability of A.
P(B) is the probability of the data (regardless of the hypothesis) [15].
Naïve Bayes classifiers are mostly used for text classification. The limitation of the Naïve Bayes model is that it treats every word in a text as independent and is equal in importance but every word cannot be treated equally important because articles and nouns are not the same when it comes to language. But due to its classification efficiency, this model is used in combination with other language processing techniques.

4.5.4.2.2 Random Forest Classifier

Random Forest classifier is a supervised ensemble algorithm. A random forest consists of multiple random decision trees. Two types of randomnesses are built into the trees. First, each tree is built on a random sample from the original data. Second, at each tree node, a subset of features is randomly selected to generate the best split [16].

Decision Tree:

The decision tree is a classification algorithm based completely on features. The tree repeatedly splits the data on a feature with the best information gain. This process continues until the information gained remains constant. Then the unknown data is evaluated feature by feature until categorized. Tree pruning techniques are used for improving accuracy and reducing the overfitting of data

4.5.4.2.3 Logistic Regression

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous [17]. The probabilities are calculated using a sigmoid function.
For example, let us take a problem where data has n features.
We need to fit a line for the given data and this line can be represented by the equation

$$z=b_0+b_1 x_1+b_2 x_2+b_3 x_3\ldots.+b_n x_n$$

here z = odds
generally, odds are calculated as
odds=p(event occurring)/p(event not occurring)

Sigmoid Function:

A sigmoid function is a special form of logistic function hence the name logistic regression. The logarithm of odds is calculated and fed into the sigmoid function to get continuous probability ranging from 0 to 1.
The logarithm of odds can be calculated by

$$\log(odds)=dot(features,coefficients)+intercept$$

and these log_odds are used in the sigmoid function to get probability.

$$h(z)=1/(1+e^{(-z)})$$

The output of the sigmoid function is an integer in the range 0 to 1 which is used to determine which class the sample belongs to. Generally, 0.5 is considered as the limit below which it is considered a NO, and 0.5 or higher will be considered a YES. But the border can be adjusted based on the requirement.

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 5, Issue 8, May 2025
Impact Factor: 7.67

4.5.4.2.4 K-Nearest Neighbors

KNN is a classification algorithm. It comes under supervised algorithms. All the data points are assumed to be in an n-dimensional space. And then based on neighbors the category of current data is determined based on the majority.
Euclidian distance is used to determine the distance between points.
The distance between 2 points is calculated as

$$d = \sqrt{(x2-x1)^2 + (y2-y1)^2}$$

The distances between the unknown point and all the others are calculated. Depending on the K provided k closest neighbors are determined. The category to which the majority of the neighbors belong is selected as the unknown data category.

4.5.4.2.5 Support Vector Machines(SVM)

It is a machine learning algorithm for classification. Decision boundaries are drawn between various categories and based on which side the point falls to the boundary the category is determined.

Support Vectors:

The vectors closer to boundaries are called support vectors/planes. If there are n categories then there will be n+1 support vectors. Instead of points, these are called vectors because they are assumed to be starting from the origin.The distance between the support vectors is called margin. We want our margin to be as wide as possible because it yields better results.

4.5.5 Experimentation

The process goes like data collection and processing then natural language processing and then vectorization then machine learning.The data is collected, cleaned, and then subjected to natural language processing techniques specified in section IV. Then the cleaned data is converted into vectors using Bag of Words and TF-IDF methods which goes like...

Accuracy:

"Accuracy is the number of correctly predicted data points out of all the data points".
The scores for Bag-of-Words and TF-IDF are visualized.

Proposed Model:

In our proposed system we combine all the models and make them into one. It takes an unknown point and feeds it into every model to get predictions. Then it takes these predictions, finds the category which was predicted by the majority of the models, and finalizes it.
To determine which model is effective we used three metrics Accuracy, Precision, and F1score. In the earlier system, we used only the F1 Score because we were not determining which model is best but which language model is best suited for classification.

4.5.6 Working Procedure

The working procedure includes the internal working and the data flow of application.
i. After running the application some procedures are automated.
1. Reading data from file
2. Cleaning the texts
3. Processing
4. Splitting the data
5. Initializing and training the models
ii. The user just needs to provide some data to classify in the area provided.

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 5, Issue 8, May 2025
Impact Factor: 7.67

iii. The provided data undergoes several procedures after submission.
1. Textual Processing
2. Feature Vector conversion
3. Entity extraction
iv. The created vectors are provided to trained models to get predictions.
v. After getting predictions the category predicted by majority will be selected.
vi. The accuracies of that prediction will be calculated
vii. The accuracies and entities extracted from the step 3 will be provided to user.
Every time the user gives something new the procedure from step 2 will be repeated.

V. RESULTS AND DISCUSSION

5.1 Language Model Selection

While selecting the best language model the data has been converted into both types of vectors and then the models been tested for to determine the best model for classifying spam.
The results from individual models are presented in the experimentation section under methodology. Now comparing the results from the models.

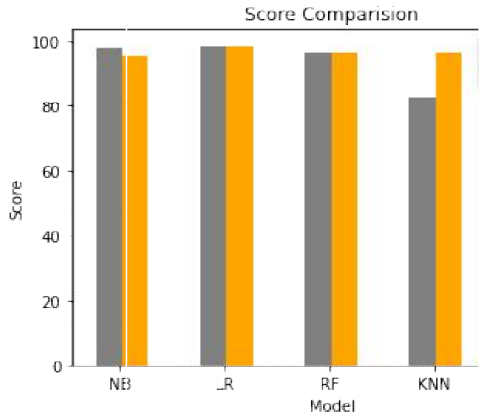


Fig no. 5.1 Bow vs TF-IDF (Cumulative)

From the figure it is clear that TF-IDF proves to be better than BoW in every model tested. Hence TF-IDF has been selected as the primary language model for textual data conversion in feature vector formation.

5.2 Proposed Model results

To determine which model is effective we used three metrics Accuracy, Precision, and F1score.
The resulted values for the proposed model are
Accuracy – 99.0
Precision – 98.5 F1 Score – 98.6

5.3 Comparison

The results from the proposed model has been compared with all the models individually in tabular form to illustrate the differences clearly.

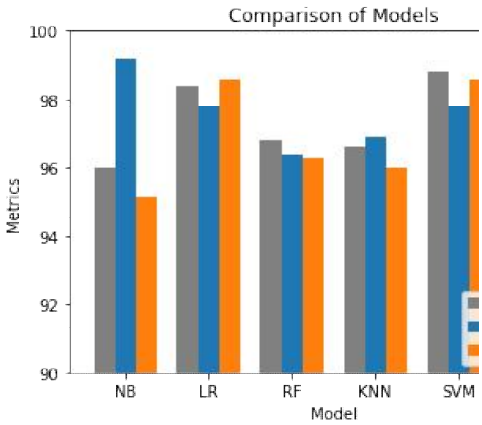
Metric Model	Accuracy	Precision	F1 Score
Naïve Bayes	96.0	99.2	95.2
Logistic Regression	98.4	97.8	98.6
Random forest	96.8	96.4	96.3
KNN	96.6	96.9	96.0

DOI: 10.48175/IJARSCT-26998

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 5, Issue 8, May 2025
Impact Factor: 7.67

SVM	98.8	97.8	98.6
Proposed model	99.0	98.5	98.6

Table no. 5.1 Models and results



The color RED indicates that the value is lower than the proposed model and GREEN indicates equal or higher.
Here we can observe that our proposed model outperforms almost every other model in every metric. Only one model (naïve Bayes) has slightly higher accuracy than our model but it is considerably lagging in other metrics.
The results are visually presented below for easier understanding and comparison

VI. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

From the results obtained we can conclude that an ensemble machine learning model is more effective in detection and classification of spam than any individual algorithms. We can also conclude that TF-IDF (term frequency inverse document frequency) language model is more effective than Bag of words model in classification of spam when combined with several algorithms. And finally we can say that spam detection can get better if machine learning algorithms are combined and tuned to needs.

6.2 Future work

There are numerous appilcations to machine learning and natural language processing and when combined they can solve some of the most troubling problems concerned with texts. This application can be scaled to intake text in bulk so that classification can be done more affectively in some public sites.
Other contexts such as negative, phishing, malicious, etc,. can be used to train the model to filter things such as public comments in various social sites. This application can be converted to online type of machine learning system and can be easily updated with latest trends of spam and other mails so that the system can adapt to new types of spam emails and texts.

REFERENCES

[1] S. H. a. M. A. T. Toma, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," in International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021.
[2] S. Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.
[3] A. L. a. S. S. S. Gadde, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, 2021.

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 5, Issue 8, May 2025
Impact Factor: 7.67

[4] V. B. a. B. K. P. Sethi, "SMS spam detection and comparison of various machine learning algorithms," in International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017.
[5] G. D. a. A. R. P. Navaney, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," in 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.
[6] S. O. Olatunji, "Extreme Learning Machines and Support Vector Machines models for email spam detection," in IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.
[7] S. S. a. N. N. Kumar, "Email Spam Detection Using Machine Learning Algorithms," in Second International Conference on Inventive Research in Computing Applications (CIRCA), 2020.
[8] R. Madan, "medium.com," [Online]. Available: https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanatio n-for-text-classification-in-nlp-with-code-8ca3912e58c3.
[9] N. D. J. a. M. M. A. M. M. RAZA, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in International Conference on Information Networking (ICOIN), 2021, 2021.
[10] A. B. S. A. a. P. M. M. Gupta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," in Eleventh International Conference on Contemporary Computing (IC3), 2018.