

**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**  
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)  
*NAAC Accredited with A++ Grade*



**Project Report**  
**on**  
**Spam & Ham Detection**

**Submitted By:**

**Atendra Rai**

**0901AM211017**

**Faculty Mentor:**

**Prof. Mir Shahnawaz Ahmad**

**Assistant Professor**

**CENTRE FOR ARTIFICIAL INTELLIGENCE**  
**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE**  
**GWALIOR - 474005 (MP) est. 1957**

**JULY-DEC. 2023**

# MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

## CERTIFICATE

This is certified that **Atendra Rai** (0901AM211017) has submitted the project report titled Spam Ham Detection under the mentorship of **Prof. Mir Shahnawaz Ahmad, Dr. Anshika Srivastava, Dr. Kritika Bansal** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Artificial Intelligence & Machine Learning** from Madhav Institute of Technology and Science, Gwalior.



**Prof. Mir Shahnawaz Ahmad**

Faculty Mentor

**Assistant Professor**

Centre for Artificial Intelligence



**Dr. R. R. Singh**

Coordinator

Centre for Artificial Intelligence

# **MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**

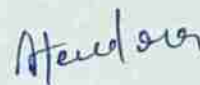
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

**NAAC Accredited with A++ Grade**

## **DECLARATION**

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in **Artificial Intelligence & Machine Learning** at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Prof. Mir Shahnawaz Ahmad, Assistant Professor**, Centre for Artificial Intelligence

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.



Atendra Rai

0901AM211017

3rd Year

Centre for Artificial Intelligence

# MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

## ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence**, for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Prof. Mir Shahnawaz Ahmad, Assistant Professor**, Centre for Artificial Intelligence, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.



Atendra Rai  
0901AM211017  
3rd Year

Centre for Artificial Intelligence

## **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.



## सार:

आजकल संचार हर चीज़ में एक प्रमुख भूमिका निभाता है चाहे वह पेशेवर हो या व्यक्तिगत। ईमेल संचार सेवा का उपयोग इसकी निःशुल्क उपयोग सेवाओं, कम लागत वाले संचालन, पहुंच और लोकप्रियता के कारण बड़े पैमाने पर किया जा रहा है। ईमेल में एक प्रमुख सुरक्षा खामी है कि कोई भी किसी को भी केवल अपनी विशिष्ट उपयोगकर्ता आईडी प्राप्त करके ईमेल भेज सकता है। इस सुरक्षा खामी का फायदा कुछ व्यवसायों और दुर्भावनापूर्ण व्यक्तियों द्वारा विज्ञापन, फिशिंग, दुर्भावनापूर्ण उद्देश्यों और अंततः धोखाधड़ी के लिए किया जा रहा है। यह एक प्रकार की ईमेल श्रेणी उत्पन्न करता है जिसे SPAM कहा जाता है। स्पैम किसी भी ऐसे ईमेल को संदर्भित करता है जिसमें विज्ञापन, असंबंधित और बार-बार आने वाले ईमेल शामिल होते हैं। इन ईमेल की संख्या दिन-ब-दिन बढ़ती जा रही है। अध्ययनों से पता चलता है कि लगभग 55 प्रतिशत ईमेल किसी न किसी प्रकार का स्पैम होता है। सेवा प्रदाताओं द्वारा इसके लिए काफी प्रयास किये जा रहे हैं। पहचान के स्पष्ट मार्करों को बदलकर स्पैम विकसित हो रहा है। इसके अलावा, सेवा प्रदाताओं की स्पैम पहचान कभी भी वर्गीकरण के साथ आक्रामक नहीं हो सकती क्योंकि इससे गलत वर्गीकरण की स्थिति में संभावित जानकारी हानि हो सकती है। इस समस्या से निपटने के लिए हम मशीन लर्निंग और प्राकृतिक भाषा प्रसंस्करण का उपयोग करके स्पैम का पता लगाने के लिए एक नई और कुशल विधि प्रस्तुत करते हैं। एक उपकरण जो स्पैम का पता लगा सकता है और वर्गीकृत कर सकता है। इसके अलावा, यह उपयोगकर्ता की सुविधा के लिए त्वरित दृश्य प्रारूप में प्रदान किए गए पाठ के बारे में जानकारी भी प्रदान करता है।

## TABLE OF CONTENTS

| CHAPTER No | TITLE                                 | PAGE No |
|------------|---------------------------------------|---------|
|            | Abstract                              | 5       |
|            | List of Figures                       | 9       |
|            | List of Tables                        | 10      |
| 1          | <b>Introduction</b>                   | 11      |
| 2          | <b>Literature Review</b>              | 13      |
|            | 2.1 introduction                      | 13      |
|            | 2.2 Related work                      | 13      |
|            | 2.3 Summary                           | 13      |
| 3          | <b>Objectives and Scope</b>           | 14      |
|            | 3.1 Problem statement                 | 14      |
|            | 3.2 Objectives                        | 14      |
|            | 3.3 Project Scope                     | 14      |
|            | 3.4 Limitations                       | 14      |
| 4.         | <b>Experimentation and Methods</b>    | 15      |
|            | 4.1 Introduction                      | 15      |
|            | 4.2 System architecture               | 15      |
|            | 4.3 Modules and Explanation           | 15      |
|            | 4.4 Requirements                      | 17      |
|            | 4.5 Workflow                          | 17      |
|            | 4.5.1 Data collection and Description | 18      |
|            | 4.5.2 Data Processing                 | 19      |
|            | 4.5.2.1 Overall Data Processing       | 19      |
|            | 4.5.2.2 Textual Data Processing       | 19      |
|            | 4.5.2.3 Feature Vector Processing     | 20      |
|            | 4.5.2.3.1 bag of words                | 20      |
|            | 4.5.2.3.2 TF-IDF                      | 20      |
|            | 4.5.3 Data Splitting                  | 24      |
|            | 4.5.4 Machine Learning                | 24      |
|            | 4.5.4.1 Introduction                  | 24      |
|            | 4.5.4.2 Algorithms                    | 24      |

|          |                                    |    |
|----------|------------------------------------|----|
|          | 4.5.4.2.1 Naïve bayes Classifier   | 24 |
|          | 4.5.4.2.2 Random Forest Classifier | 24 |
|          | 4.5.4.2.3 Logistic Regression      | 24 |
|          | 4.5.4.2.4 K-Nearest Neighbors      | 24 |
|          | 4.5.4.2.5 Support Vector machines  | 24 |
|          | 4.5.5 Experimentation              | 25 |
|          | 4.5.6 User Interface(UI)           | 26 |
|          | 4.5.7 Working Procedure            | 27 |
| <b>5</b> | <b>Results and Discussion</b>      | 30 |
|          | 5.1 Language Model selection       | 30 |
|          | 5.2 Proposed Model                 | 31 |
|          | 5.3 Comparison                     | 32 |
|          | 5.4 Summary                        | 32 |
| <b>6</b> | <b>Conclusion and Future Scope</b> | 33 |
|          | 6.1 Conclusion                     | 33 |
|          | 6.2 Future Work                    | 33 |
|          | <b>References</b>                  | 34 |



## List of Figures

| Fig No | Title                              | Pg no |
|--------|------------------------------------|-------|
| 4.1    | Architecture                       | 15    |
| 4.2    | Workflow                           | 17    |
| 4.3    | Enron Data                         | 18    |
| 4.4    | Ling spam                          | 18    |
| 4.5    | Naïve Bayes(Bow vs TF-IDF)         | 27    |
| 4.6    | Logistic Regression(Bow vs TF-IDF) | 28    |
| 4.7    | Neighbors vs Accuracy(KNN)         | 28    |
| 4.8    | KNN(Bow vs TF-IDF)                 | 29    |
| 4.9    | Random Forest(trees vs scores)     | 29    |
| 4.10   | Random Forest(Bow vs TF-IDF)       | 29    |
| 4.11   | SVM(Bow vs TF-IDF)                 | 30    |
| 5.1    | Bow vs TF-IDF(Cumulative)          | 32    |
| 5.2    | Comparision of Models              | 33    |

### List of Tables

| Table number | Table Name                 | Page no |
|--------------|----------------------------|---------|
| 4.1          | Term Frequency             | 22      |
| 4.2          | Inverse document frequency | 22      |
| 4.3          | TF-IDF                     | 22      |
| 5.1          | Models and results         | 33      |

# 1. 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. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary.

Text classification is important to structure the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. Machine learning can make more accurate precisions in real-time and help to improve the manual slow process to much better and faster analysing big data. It is important especially to a company to analyse text data, help inform business decisions and even automate business processes.

In this project, machine learning techniques are used to detect the spam message of a mail. Machine learning is where computers can learn to do something

without the need to explicitly program them for the task.

It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio.

A combination of algorithms are used to learn the classification rules from messages. These algorithms are used for classification of objects of different classes. These algorithms are provided with pre labelled data and an unknown text. After learning from the prelabelled data each of these algorithms predict which class the unknown text may belong to and the category predicted by majority is considered as final.



## **2. 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, analysing, summarising and evaluating the reading materials based on the project.

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]

### **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.

## **3. 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 an 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, process and make use of the dataset for training and testing the model created.
- iii. Analyse 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 analyse the message

## 4. 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.

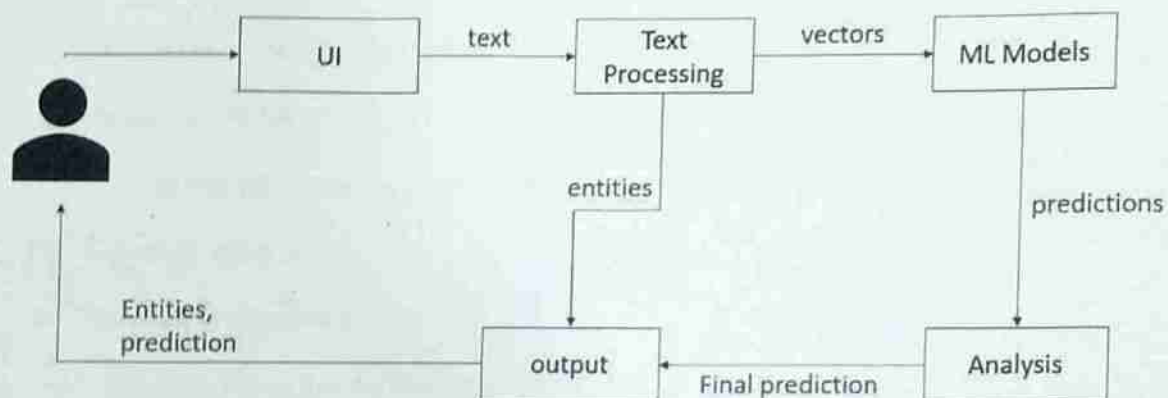


fig no. 4.1 Architecture

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.

A more complicated and detailed view of architecture is presented in the workflow section.

### 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

## **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

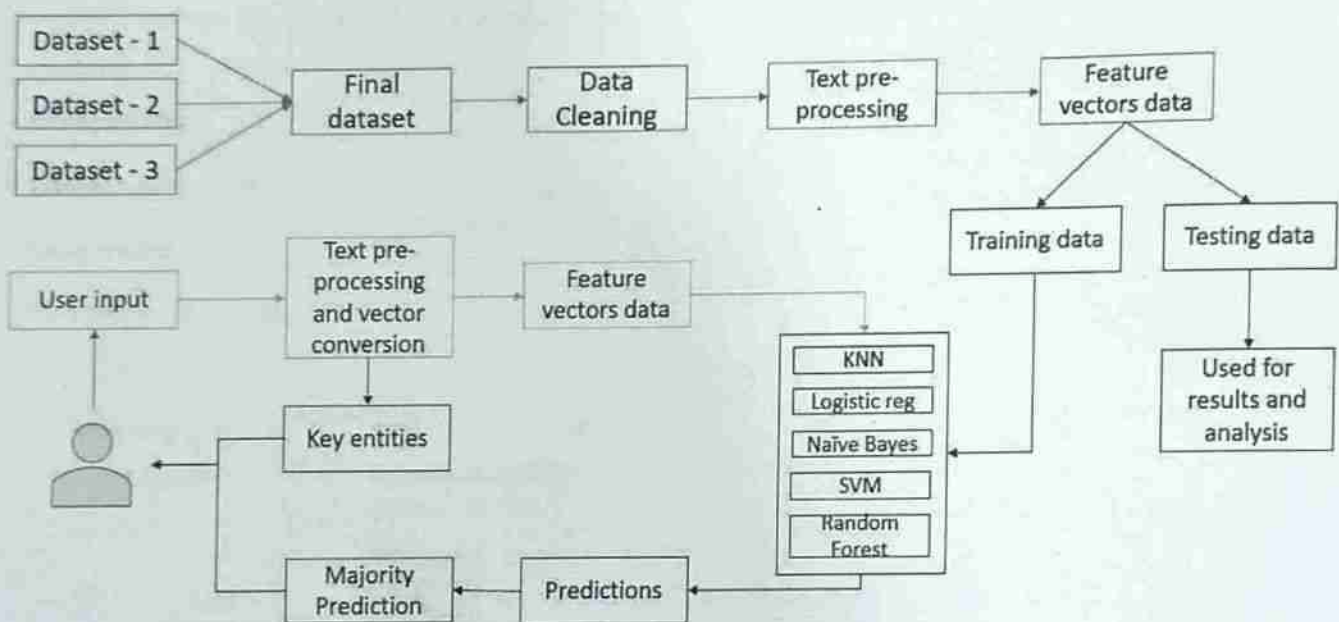
Code Editor – Pycharm, VS Code, Built in IDE

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.

## 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
- For the purpose of accuracy and diversity in data multiple datasets are taken.

2 datasets containing approximately over 12000 mails and their labels are used for training and testing the application.

- 6000 spam mails are taken for generalisation of data and to increase the accuracy.

### 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

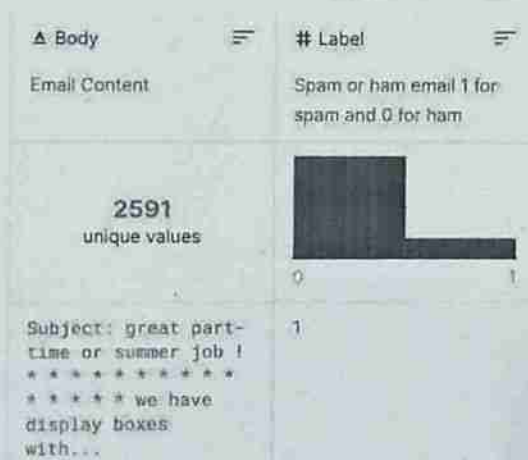


fig no. 4.3 enron spam

**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

| Body   | # Label   |
|--|---|
| Email Content  | Spam or ham email 1 for spam and 0 for ham  |
| 9687<br>unique values  |  |
| Subject: stock promo<br>mover : cwid * * *<br>urgent investor<br>trading alert * * *<br>weekly stock pick -<br>* 2/1/0 | 1   |

fig no. 4.4 lingspam

## 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).

Data cleaning, Data Merging these procedures are completely done using Pandas library.

### 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).

This process is done using NLTK pre-processing library of python.

- **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

This process gives key-words through entity extraction.

This process is done using chunking in regex and NLTK lemmatization.

- **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.

Then all the sentences are combined to form a text. While the overall data processing is done only to datasets, the textual processing is done to both training data, testing data and also user input data.



### 4.5.2.3 Feature Vector Formation

- The texts are converted into feature vectors(numerical data) using the words present in all the texts combined
- This process is done using countvectorization of **NLTK** library.
- The feature vectors can be formed using two language models Bag of Words and Term Frequency-inverse Document Frequency.

#### 4.5.2.3.1 Bag of Words

Bag of words is a language model used mainly in text classification. A bag of words represents the text in a numerical form.

The two things required for Bag of Words are

- A vocabulary of words known to us.
- A way to measure the presence of words.

Ex: a few lines from the book "A Tale of Two Cities" by Charles Dickens.

**" It was the best of times, it  
was the worst of times, it  
was the age of wisdom,  
it was the age of foolishness, "**

The unique words here (ignoring case and punctuation) are:

[ "it", "was", "the", "best", "of", "times", "worst", "age", "wisdom", "foolishness" ]

The next step is scoring words present in every document.

After scoring the four lines from the above stanza can be represented in vector form as "It was the best of times" = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

"it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0] "it was

the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0] "it was the age

of foolishness"= [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

This is the main process behind the bag of words but in reality the vocabulary even from a couple of documents is very large and words repeating frequently and important in nature are taken and remaining are removed during the text processing stage.

#### 4.5.2.3.2 Term Frequency-inverse document frequency

Term frequency-inverse document frequency of a word is a measurement of the importance of a word. It compares the repentence of words to the collection of documentsand calculates the score.

Terminology for the below formulae:

t – term(word)

d – document(set of words)

N – count of documents

The TF-IDF process consists of various activities listed below.

**i) Term Frequency**

The count of appearance of a particular word in a document is called term frequency ( $tf_t$ ) =

**ii) Document Frequency**

Document frequency is the count of documents the word was detected in. We consider one instance of a word and it doesn't matter if the word is present multiple times. ( $df_t$ ) =

**iii) Inverse Document Frequency**

- IDF is the inverse of document frequency.
- It measures the importance of a term t considering the information it contributes. Every term is considered equally important but certain terms such as (are, if, a, be, that, ..) provide little information about the document. The inverse document frequency factor reduces the importance of words/terms that has high recurrence and increases the importance of words/terms that are rare.

$$idf_t = 1 / df_t$$

Finally, the TF-IDF can be calculated by combining the term frequency and inverse document frequency.

$$tfidf_t = tf_t * idf_t$$

the process can be explained using the following example:

**Document 1 It is going to rain today.**

**Document 2 Today I am not going outside.**

**Document 3 I am going to watch the season premiere."**

The Bag of words of the above sentences is [going:3, to:2, today:2, i:2, am:2, it:1, is:1, rain:1]

Then finding the term frequency

table no. 4.1 Term frequency

| Words | IDF Value   |
|-------|-------------|
| Going | $\log(3/3)$ |
| To    | $\log(3/2)$ |
| Today | $\log(3/2)$ |
| I     | $\log(3/2)$ |
| Am    | $\log(3/2)$ |
| It    | $\log(3/1)$ |
| Is    | $\log(3/1)$ |
| rain  | $\log(3/1)$ |

Then finding the inverse document frequency

table no. 4.2 inverse document frequency

| Words | Document1 | Document2 | Document3 |
|-------|-----------|-----------|-----------|
| Going | 0.16      | 0.16      | 0.12      |
| To    | 0.16      | 0         | 0.12      |
| Today | 0.16      | 0.16      | 0         |
| I     | 0         | 0.16      | 0.12      |
| Am    | 0         | 0.16      | 0.12      |
| It    | 0.16      | 0         | 0         |
| Is    | 0.16      | 0         | 0         |
| rain  | 0.16      | 0         | 0         |

Applying the final equation the values of tf-idf becomes

| Words/<br>documents | going | to   | Today | i    | am   | if   | it   | rain |
|---------------------|-------|------|-------|------|------|------|------|------|
| Document1           | 0     | 0.07 | 0.07  | 0    | 0    | 0.17 | 0.17 | 0.17 |
| Document2           | 0     | 0    | 0.07  | 0.07 | 0.07 | 0    | 0    | 0    |
| Document3           | 0     | 0.05 | 0     | 0.05 | 0.05 | 0    | 0    | 0    |

table no. 4.3 TF-IDF

Using the above two language models the complete data has been converted into two kinds of vectors and stored into a csv type file for easy access and minimal processing.

### **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 analyse 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 countvectorization and TF-IDF stages in the workflow the data is converted into vector form(numerical form) which is used for training and testing models.

For our study various machine learning models are compared to determine which method is more suitable for this task. The models used for the study include Logistic Regression, Naïve Bayes, Random Forest Classifier, K Nearest Neighbors, and Support Vector Machine Classifier and a proposed model which was created using an ensemble approach.

#### **4.5.4.2 Algorithms**

a combination of 5 algorithms are used for the classifications.

**4.5.4.2.1 Naïve Bayes Classifier**

**4.5.4.2.2 Random Forest Classifier**

**4.5.4.2.3 Logistic Regression**

**4.5.4.2.4 K-Nearest Neighbors**

**4.5.4.2.5 Support Vector Machines(SVM)**



## 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...

The Data is split into Training data and Testing Data in an 80-20 split ratio. The training and testing data is converted into Bag-of-Words vectors and TF-IDF vectors.

There are several metrics to evaluate the models but accuracy is considered for comparing BoW and TF-IDF models. Accuracy is generally used to determine the efficiency of a model.

### Accuracy:

"Accuracy is the number of correctly predicted data points out of all the data points".

### Naïve Bayes Classification algorithm:

Two models, one for Bow and one for TF-IDF are created and trained using respective training vectors and training labels. Then the respective testing vectors and labels are used to get the score for the model.

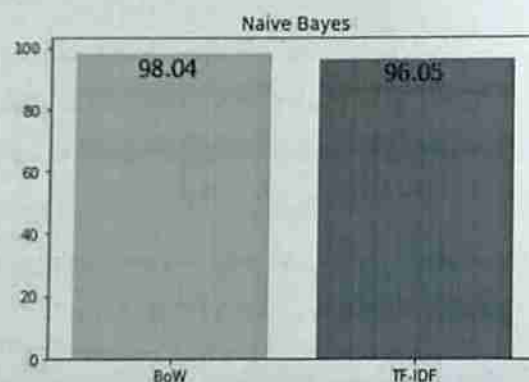


fig no. 4.5 naïve Bayes

The scores for Bag-of-Words and TF-IDF are visualized.

The scores for the Bow model and TF-IDF models are 98.04 and 96.05 respectively for using the naïve bayes model.

### Logistic Regression:

Two models are created following the same procedure used for naïve Bayes models and then tested the results obtained are visualized below.

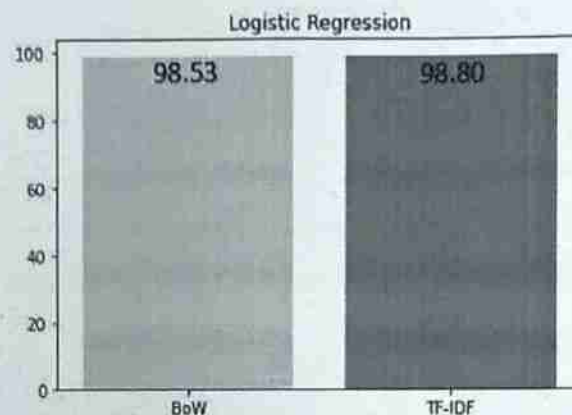


fig no. 4.6 Logistic Regression (Bow vs TF-IDF)

The scores for BoW and TF-IDF models are 98.53 and 98.80 respectively.

### K-Nearest Neighbors:

Similar to the above models the models are created and trained using respective vectors and labels. But in addition to the data, the number of neighbors to be considered should also be provided.

Using Iterative Method  $K=3$  (no of Neighbors) provided the best results for the BoW model and  $K=9$  provided the best results for the TF-IDF model.

Using the K values the scores for BOW and TF-IDF are visualized below.

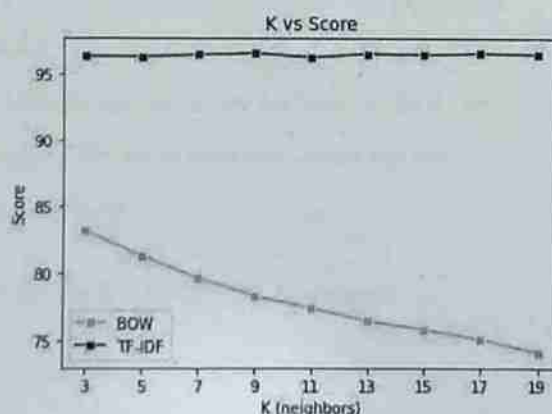


fig no. 4.7 Neighbors vs Accuracy

respectively the scores are calculated and are presented below.

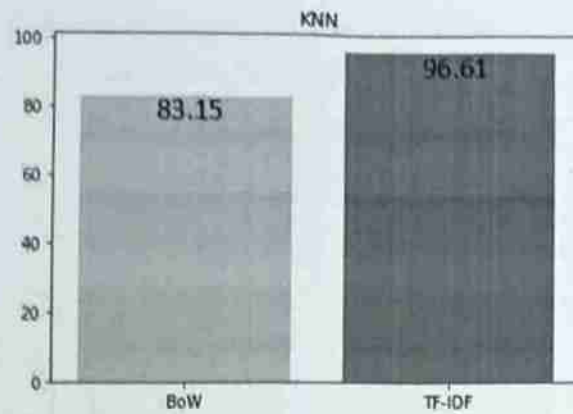


fig no. 4.8 KNN (Bow vs TF-IDF)

### Random Forest:

Similar to previous algorithms two models are created and trained using respective training vectors and training labels. But the number of trees to be used for forest has to be provided.

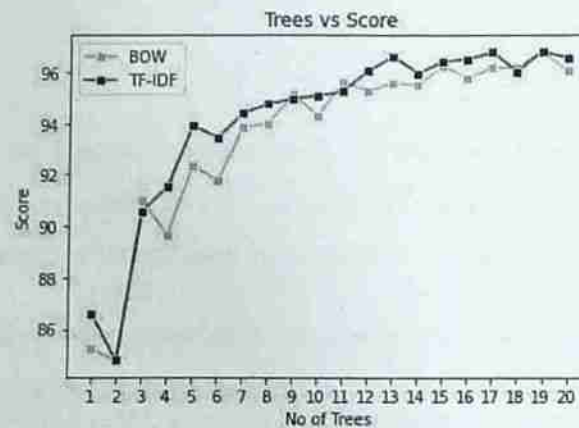
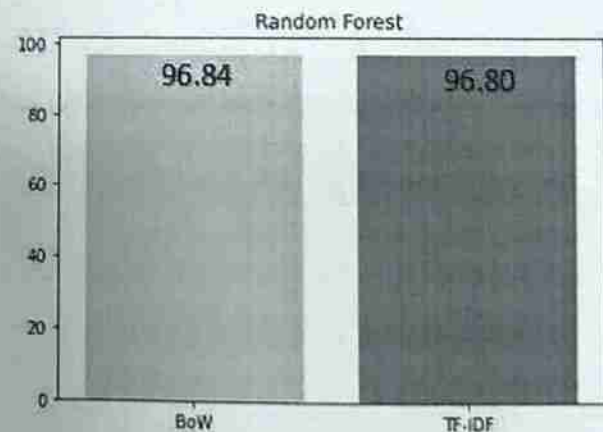


fig no. 4.9 Random

### Forest (trees vs score)

Using the Iterative method best value for the number of trees is determined. From the results, it is clear that 19 estimators provide the best score for both the BoW and TF-IDF models. The no of trees and scores for both models are visualized.

The scores for BoW and TF-IDF models are visualized.



( fig no. 4.10 Random Forest(bow vs tfidf)

### Support Vector Machines (SVM):

Finally, two SVM models, one for BoW and one for TF-IDF are created and then trained using respective training vectors and labels. Then tested using testing vectors and labels.

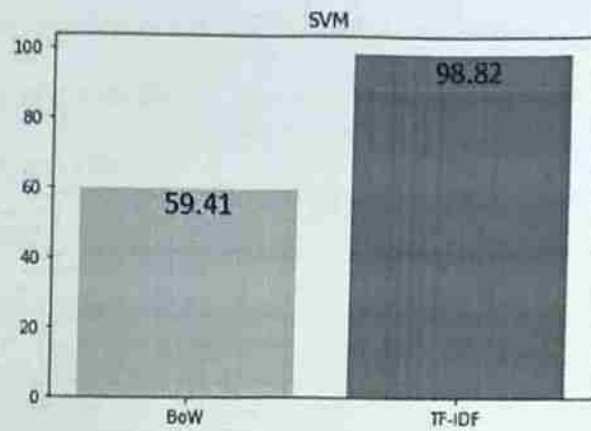


fig no. 4.11 SVM(Bow vs TF\_IDF)

The scores for BoW and TF-IDF models are 59.41 and 98.82 respectively.

### 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 User Interface(UI)

interface (UI) is an important component in this application. The user only interacts with the interface.

The UI of this project has been constructed with the help of an open source library called streamlit.

The complete information and API reference sheet can be obtained from



### 4.5.7 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. Initialising and training the models
- ii. The user just needs to provide some data to classify in the area provided.
- 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.

## 5. 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.

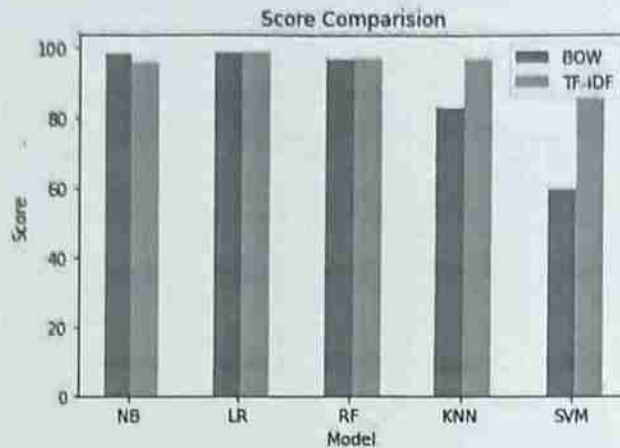


fig no. 5.1 Bow vs TF-IDF (Cumulative)

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     |
| SVM                 | 98.8     | 97.8      | 98.6     |
| Proposed model      | 99.0     | 98.5      | 98.6     |

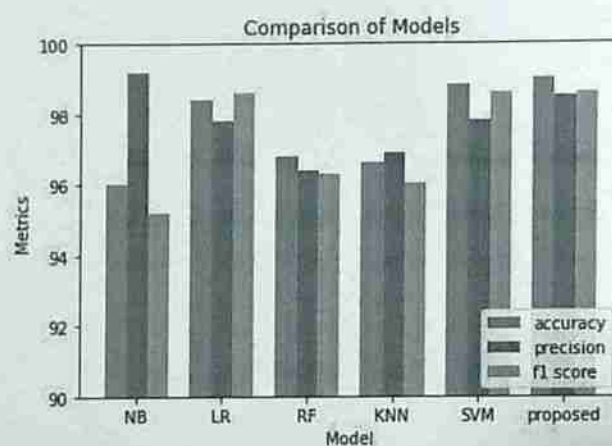


Table no. 5.1

The color RED value is lower than model and GREEN higher.

Models and results

indicates that the the proposed indicates equal or

fig no. 5.2 Comparision of Models

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.

From the above comparison barchart we can clearly see that all models individually are not as efficient as the proposed method.

## 5.4 Summary

There are two main tasks in the project implementation. Language model selection for completing the textual processing phase and proposed model creation using the individual algorithms. These two tasks require comparison from other models and select of various parameters for better efficiency.

During the language model selection phase two models, Bag of Words and TF-IDF are compared to select the best model and from the results obtained it is evident that TF-IDF performs better.

During the proposed model design various algorithms are tested with different parameters to get best parameters. Models are merged to form a ensemble algorithm and the results obtained are presented and compared above. It is clear from the results that the proposed model outperforms others in almost every metric derived.



## **6. 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 applications 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.
- [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.