Maharshi Karve Stree Shikshan Samstha's
Cummins College of Engineering for Women, Pune
(An Autonomous Institute affiliated to Savitribai Phule Pune University)

SPAM VS HAM EMAIL CLASSIFIER

Semester : AY 2025-26 , Sem I

Department : Computer Engineering

Presented by

UCE2023404-Riya Handu

UCE2023476-Gauri Warme

UCE2024006-Samruddhi Adhikary

OVERVIEW

- Introduction
- Problem
- Objectives
- ML models for classification
- Methodology
- Experimental Setup
- Result
- Discussion
- Conclusion
- Future Scope
- Reference

INTRODUCTION

- Email is a widely used communication medium, but spam emails and sms have become a major threat.
- Spam includes phishing links, malware, financial scams, and harmful URLs.
- Nearly half of daily global email traffic is spam, requiring strong filtering systems.
- Traditional rule-based filters rely on keywords and blacklists but fail against evolving attacks.
- Machine learning adapts better by learning patterns from historical email and sms data.
- This study focuses on ML-based text classification models using TF-IDF features.
- Four commonly used algorithms are evaluated: Naïve Bayes, Logistic Regression, KNN, and Random Forest.

PROBLEM

- Users receive a large number of unwanted and harmful emails every day.
- These spam messages often contain scams, fake offers, phishing links, or malware.
- Traditional rule-based filters fail because attackers keep changing the way spam looks.
- Keyword-based systems cannot detect new or modified spam patterns.
- Many safe emails get marked as spam, and many spam emails enter the inbox.
- There is a need for a machine-learning model that can learn from data and classify emails accurately.

OBJECTIVE

- To build a system that automatically identifies spam emails and sms.
- To protect users from scams, phishing links, and unwanted messages using NLP.
- To clean and analyze email text using machine learning.
- To compare different ML models and choose the most accurate one.
- To provide a simple Streamlit interface for real-time spam detection.

ML MODELS FOR CLASSIFICATION

Naïve Bayes

- From the study by Sultana & Thashina, we found that NB is widely used in early spam filters due to its simplicity and fast computation.
- The paper shows NB performs well when spam contains repetitive keywords like “free,” “win,” “urgent.”

Logistic Regression

- Alsuwit et al. (2024) report that Logistic Regression is one of the most reliable linear models for large-scale spam detection.
- Their findings indicate that LR works particularly well with TF-IDF features and high-dimensional datasets.

k-Nearest Neighbors (k-NN)

- The review by Tusher et al. (2024) discusses the use of k-NN in earlier spam detection research.
- Their analysis shows that while k-NN is simple and intuitive, its performance drops in high-dimensional text spaces because distance metrics become less meaningful.

ML MODELS FOR CLASSIFICATION

Random Forest

- According to Speiser et al. (2019) and the comparative work by Alsuwit et al. (2024), Random Forest provides better generalization than a single decision tree.
- These studies emphasize that RF reduces overfitting through bagging and random feature selection, making it effective for text classification.

TF-IDF & Feature Extraction

- Shafi'i et al. (2018) and Hegde (2021) highlight preprocessing methods such as tokenization, stopword removal, and stemming as essential for improving spam detection accuracy.
- Their work supports TF-IDF as the most widely used and effective technique for converting text into discriminative numerical features.

DATASET AND TOOLS

Dataset

- **Kaggle** : Email Spam vs Ham Dataset
- Contains labelled messages: spam or ham

- Used for training, testing, and evaluating the ML models (80% - train , 20% test)

- The dataset contains 20k + email rows with 3 main columns:

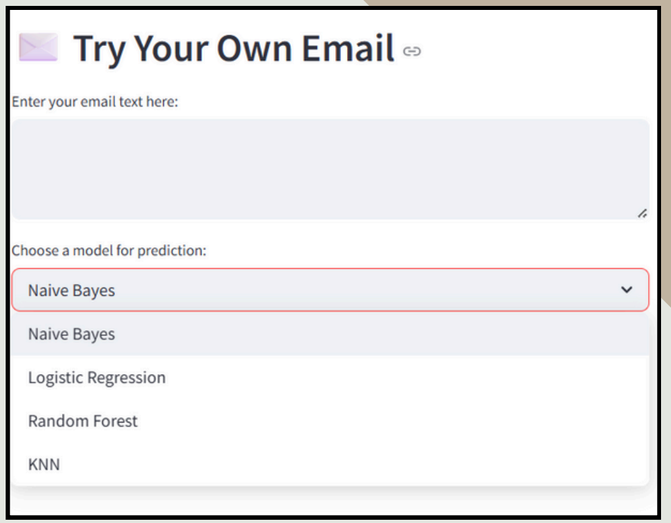
1. label → indicates whether the text is spam or ham
2. text → the actual email/message content
3. type → the text is email or SMS.

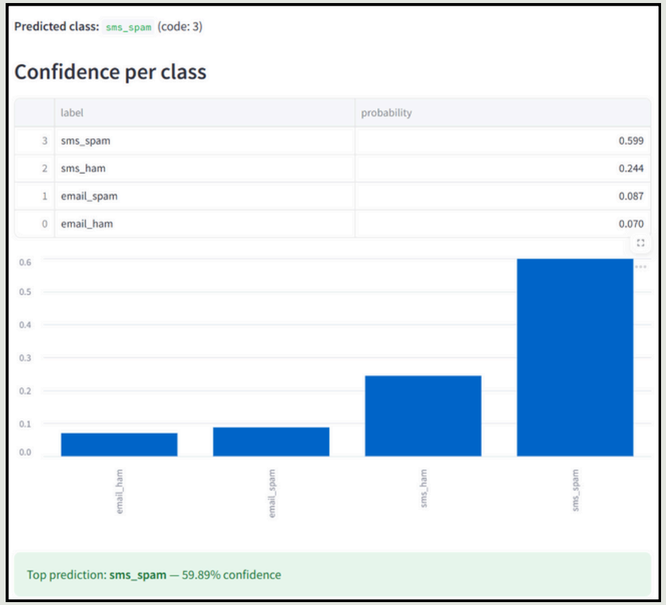
Technologies

- **Python** : main programming language
- **Streamlit** : for building the interactive web app
- **scikit-learn** : machine learning algorithms & metrics
- **Pandas / NumPy :** data handling and processing
- **Matplotlib / Seaborn** : for visualizations
- **sklearn ,re and String** for preprocessing
- **TF-IDF (Sklearn) :** for converting text into numerical features
- **NLTK:** For detecting phishing links and domain,etc.
- **Joblib** : to save and load trained models

KEY FEATURES

- Automatic spam vs ham classification for both ham and spam using machine learning.
- Clean and preprocess email and sms text using TF-IDF–based feature extraction and NLP based preprocessing.
- Supports multiple ML models: Naïve Bayes, Logistic Regression, Random Forest, and KNN.
- User can select the model they want to test or use for prediction.
- Real-time 4 way classification and prediction through an easy and interactive Streamlit interface.
- Visual performance comparison using confusion matrices .





METHODOLOGY

- The system starts by using the Kaggle Spam vs Ham dataset (combined) as the labelled source for email messages.

- The text is cleaned using preprocessing steps such as :
  1.Lowercasing
  2. Removal of digits
  3. Removing punctuation
  4. Tokenization on whitespace
  5. Removing English stop words
  6. Rejoining tokens

- Label encoding

- The cleaned messages are then transformed into numerical vectors using a 3000-feature TF-IDF representation.

NLP:

- Lemmatizes words to their base form. (running –> run)
- Adds security features (URLs, suspicious domains, phishing keywords, all-caps, exclamations)
- Helps the model detect spam/ham patterns more accurately

- These features are used to train multiple models including Naïve Bayes, Logistic Regression, Random Forest, and KNN.

- Finally, the models are evaluated using accuracy, precision, recall, F1-score, and ROC curves to identify the best-performing classifier for deployment.

PRE-PROCESSING RESULTS

Cleaning Text Data...

Sample Cleaned Text (First 5 Rows)

	text	clean_text
0	WINNER!!! You have won $1000 cash!!! Click here NOW... .	winner won cash click
1	Hey, r u coming 2day? :) meeting @ 5pm!!	hey r u coming day meeting pm
2	URGENT!! Your account will be BLOCKED. Visit http://xyz.com/secure	urgent account blocked visit httpxyzcomsecure
3	FREE entry in 2 a wkly draw!! Txt WIN to 80085!!!?	free entry wkly draw txt win
4	Hello John, Your invoice #INV-9987 is due on 12/09/24.	hello john invoice inv

✅ Text cleaning complete!

TF-IDF WORKING

$$TFIDF(T,D) = TF(T,D) \times IDF(T)$$

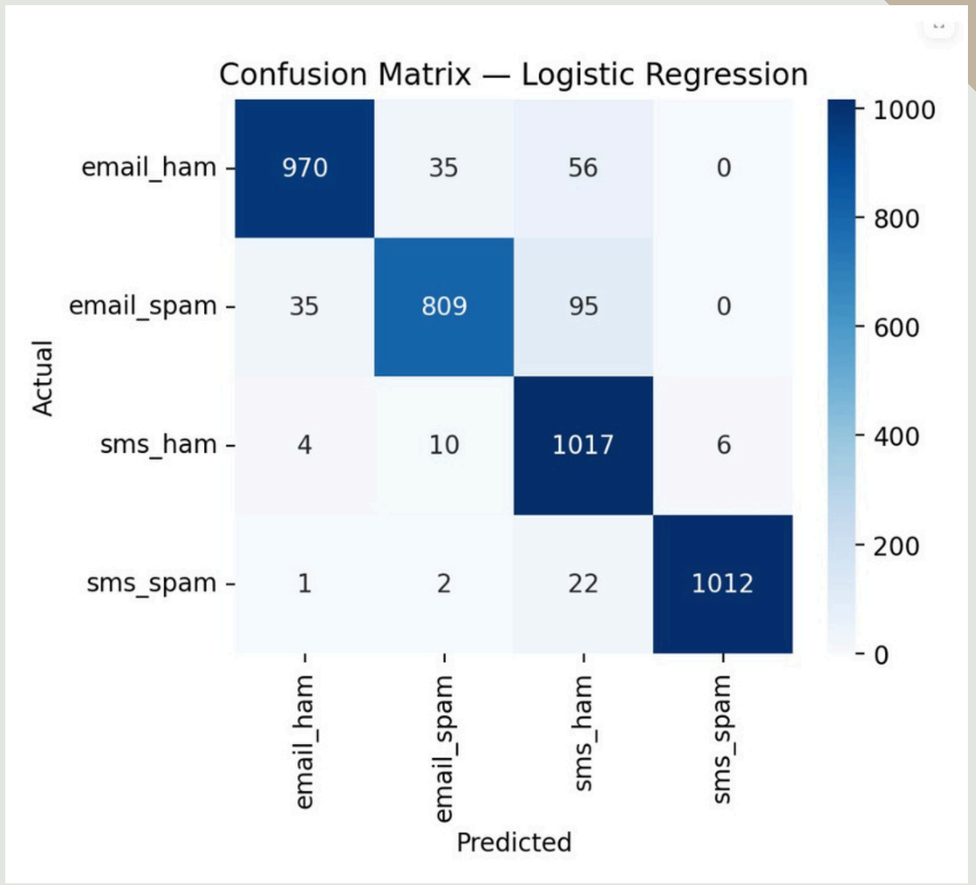
	text	tf	idf
0	Eddard Stark is a king in the north.	1	3
1	A king but one king : kings are everywhere.	2	3
2	Hodor was different : he was not a king .	1	3
3	But the North could not change without him.	0	3

	king	was	the	not	a		he	one	north	kings	is	in	him	everywhere	A	different	could	change	but	are	Stark	North	Hodor	Eddard
0	0.333333	0.0	0.5	0.0	0.5	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.666667	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
2	0.333333	2.0	0.0	0.5	0.5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	0.000000	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0

RESULTS

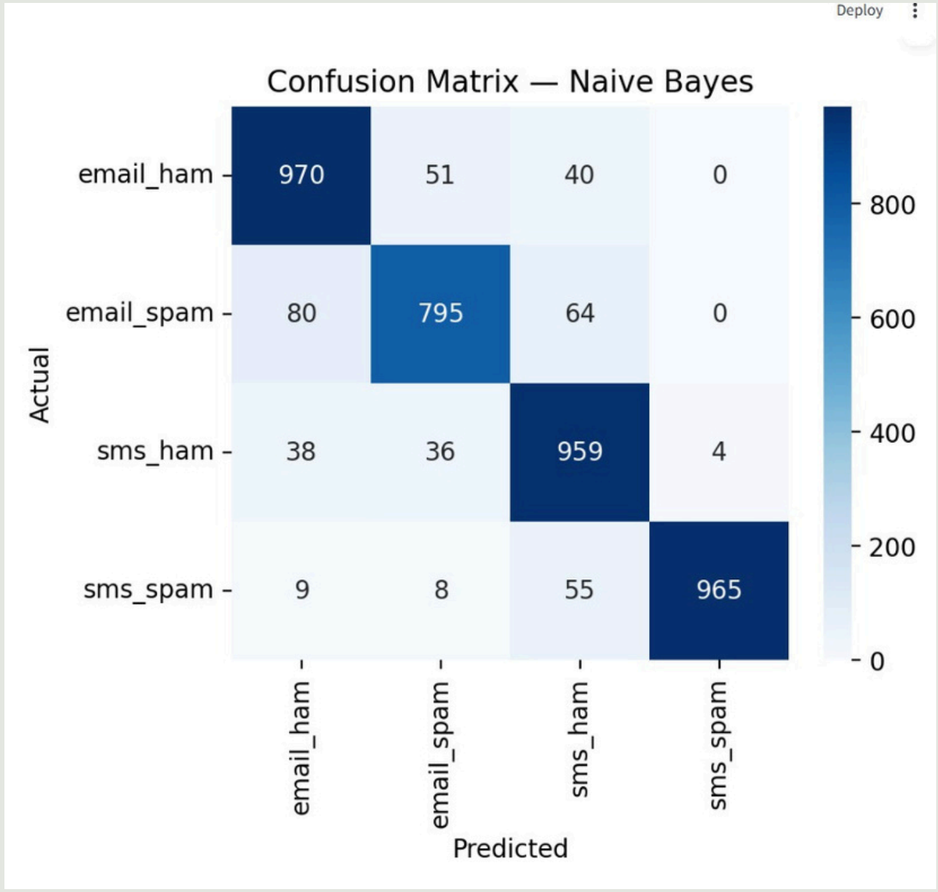
Logistic Regression

- Logistic Regression achieved the highest accuracy among all models.
- It performed strongly across precision, recall, and F1-score.
- It produced very few false negatives and kept false positives low.
- TF-IDF features helped it form clear and stable decision boundaries.
- Overall, it proved to be the most reliable model for spam detection.



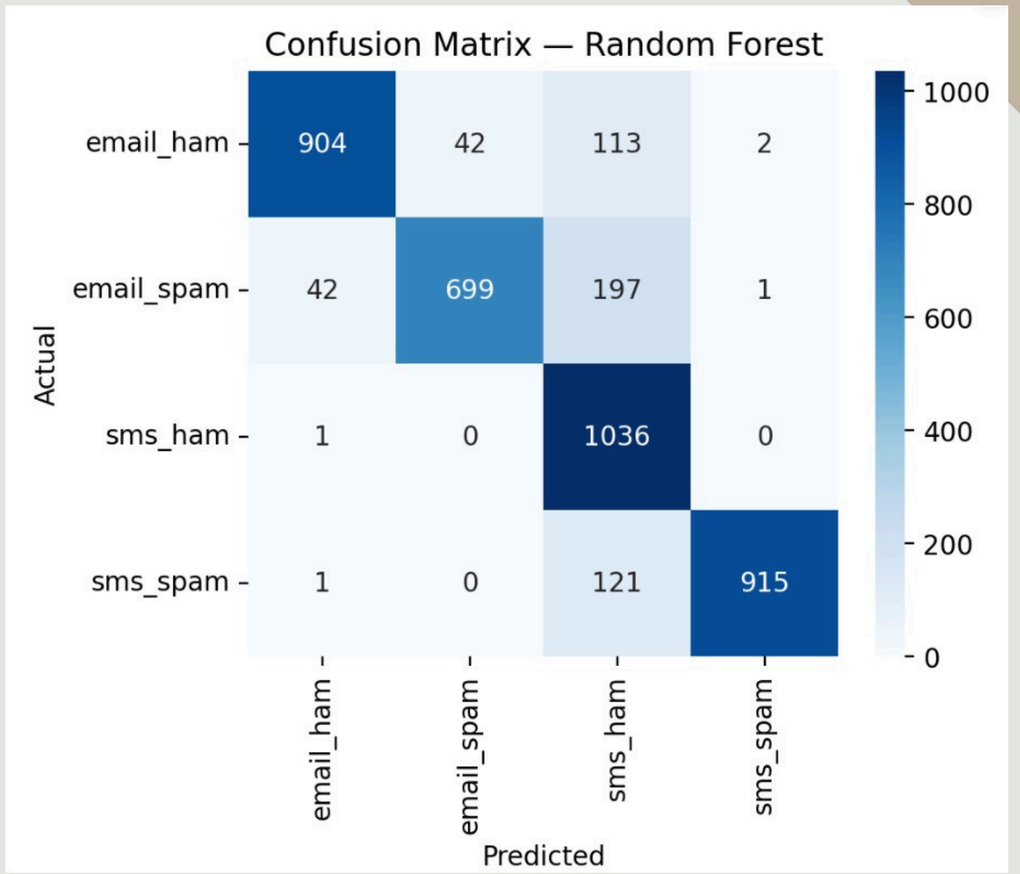
Naive Bayes

- Showed strong overall accuracy with very high recall for spam emails.
- Confusion matrix indicated a moderate number of false negatives but relatively low false positives.
- Performed well on large, sparse TF-IDF features due to its probabilistic nature.
- Very fast training and prediction, making it efficient for real-time filtering.
- Slightly struggled with emails containing mixed or informal language.



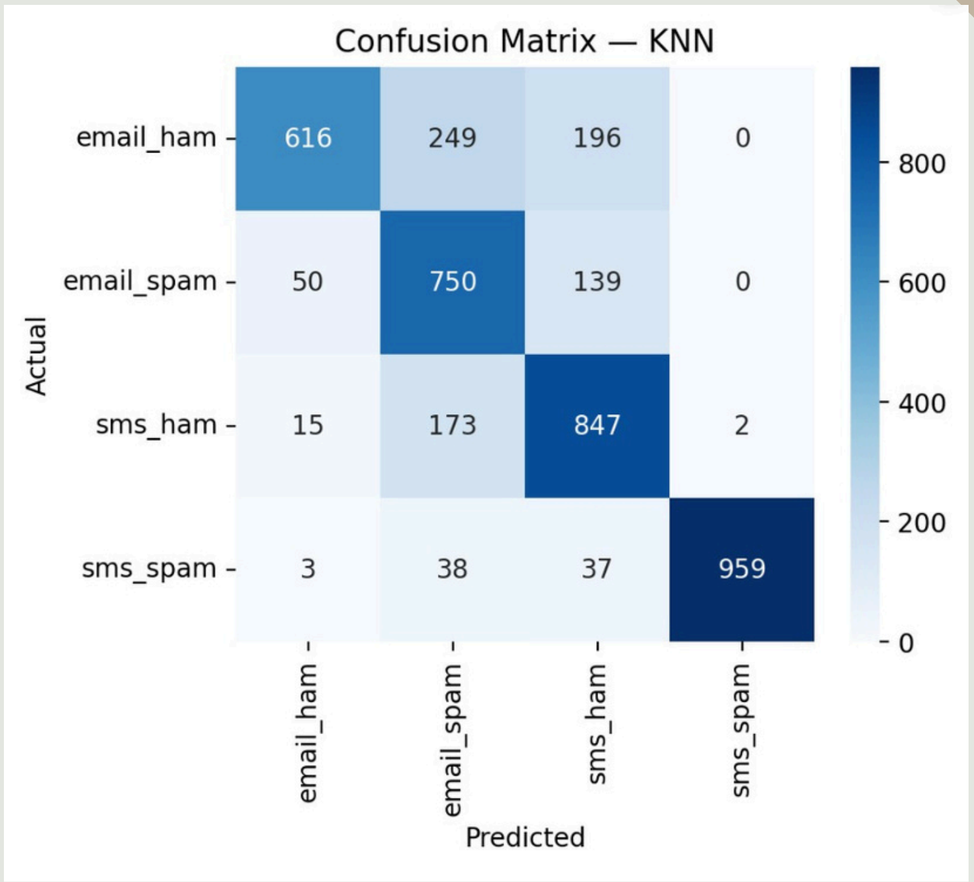
Random Forest

- Delivered strong accuracy with particularly high recall for spam messages.
- Confusion matrix revealed more false positives than Logistic Regression, showing slight over-flagging of spam.
- Benefited from ensemble voting, capturing nonlinear patterns in the dataset.
- Performed reliably but required more computational resources than linear models.
- Showed robustness to noise and uneven feature distributions.

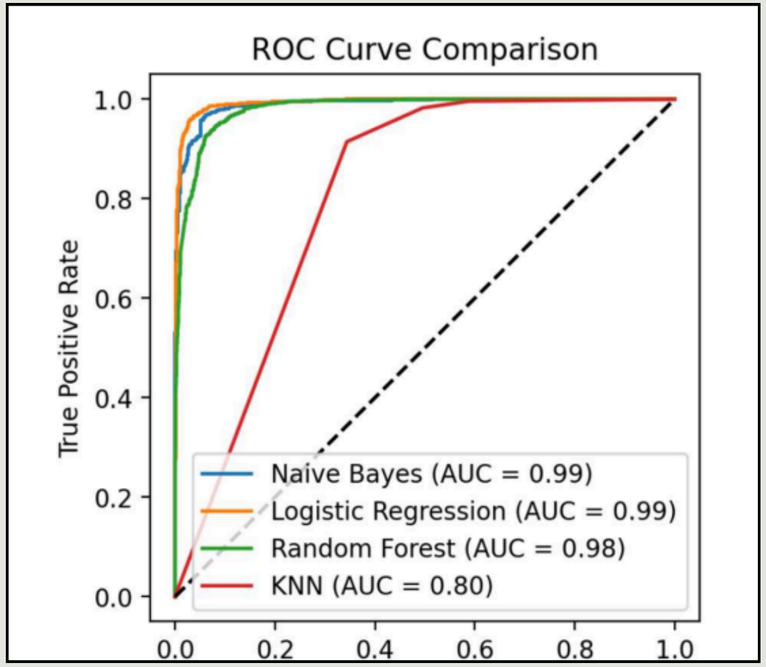


K-Nearest Neighbors (k-NN)

- Showed the weakest performance among all models on the heatmap, with many false positives.
- Struggled with high-dimensional TF-IDF vectors, leading to poor separation between classes.
- Predictions were slow due to distance calculations during inference.
- Performed inconsistently on sparse data, misclassifying many legitimate emails as spam.
- Served as a useful baseline but not suitable for production-level spam filtering.



- **Logistic Regression:** achieved the highest accuracy and F1-score is (0.934) with an AUC of 0.99, showing excellent discrimination between spam and ham.
- **Naïve Bayes:** also performed strongly, with accuracy (0.905) and AUC (0.99), indicating high reliability across all evaluations.
- **Random Forest:** reached accuracy (0.875) and AUC (0.98), handling high-dimensional text well but slightly below the linear models.
- **KNN:** had the lowest accuracy (0.778) and AUC (0.80), confirming its weaker ability to separate classes in complex textual data.



Model performance

	Model	Accuracy	Precision	Recall	F1 Score
0	Naive Bayes	0.905500	0.908000	0.905500	0.905900
1	Logistic Regression	0.934700	0.938500	0.934700	0.934900
2	Random Forest	0.872400	0.899200	0.872400	0.875100
3	KNN	0.778600	0.808300	0.778600	0.780200

Overall

Logistic Regression proves to be the most reliable and effective model for spam detection.

FUTURE SCOPE

- Improve accuracy using hybrid/ensemble models combining NB, LR, and RF.
- Apply feature selection methods (Chi-square, Mutual Information) to reduce noise and speed up training.
- Automatic integration in the user’s email and sms app

CONCLUSION

The system accurately classifies both emails and SMS messages into spam and ham categories using NLP, TF-IDF, and engineered security features. Logistic Regression and Random Forest gave the best overall performance. The project proves that a well-designed ML pipeline can reliably detect spam across multiple communication channels without needing deep learning.

REFERENCE

M. H. Alsuwit, M. A. Haq, and M. A. Aleisa, “Advancing Email Spam Classification using Machine Learning and Deep Learning Techniques”, Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 14994– 15001, Aug. 2024

https://scikit-learn.org/1.4/tutorial/text_analytics/working_with_text_data.html scikit-learn

Sultana, Thashina. (2020). Email based Spam Detection. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS060087Y.

M. A. Shafi’i, S. Maryam, O. Oluwafemi, I. Ismaila, and K. A. John, “Comparative Analysis of Classification Algorithms for Email Spam Detection,” I. J. Computer Network and Information Security, Jan. 2018.

J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, ‘‘A comparison of random forest variable selection methods for classification prediction modeling,’’ Expert Syst. Appl., vol. 134, pp. 93–101, Nov. 2019.

E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi and M. Uddin, "Email Spam: A Comprehensive Review of Optimize Detection Methods, Challenges, and Open Research Problems," in IEEE Access, vol. 12, pp. 143627-143657, 2024, doi: 10.1109/ACCESS.2024.3467996.

Thank You

For your attention