INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Spam Detection Using Machine Learning

TF-IDF & Logistic Regression

Done By: Anuradha K

About me

- **Background ? (B-tech or M-tech)**

B.Arch Graduate

- **Why you want to learn Data Science**

I enjoy learning new things, and data science is a field that's always growing and changing. I want to add new skills like data analysis, programming, and machine learning to my skill set.
In architecture, we often make choices based on design principles. I'm excited to learn how to make decisions based on data, which can provide more precise and impactful results.

- **Any work experience**

No work experience.

- **Share your linkedin and github profile urls**

Anuradha Kilaparthi | LinkedIn
anuradhak0801 (Anuradha K) (github.com)

INNOMATICS RESEARCH LABS

PROBLEM STATEMENT AND USE CASE DOMAIN

Problem Statement:

Every day, people receive tons of unwanted spam messages — from fake offers to harmful links. These messages can be annoying, misleading, and even dangerous.

The main problem that we can solve is: How can we automatically detect and filter out these spam messages using machine learning?

A smart spam detection system saves time, improves security, and enhances user experience.

Use Case Domain:

This project belongs to the domain of Natural Language Processing (NLP) and Text Classification.

NLP is a branch of AI that deals with understanding and working with human language.
In this case, we teach a machine to read and understand messages, and then decide whether a message is spam or not spam (ham).
Spam detection is widely used in Email filtering, SMS spam protection, Chatbot and social media moderation etc.

OBJECTIVE

The goal of this project is to:

- Build a machine learning model that can accurately classify messages as spam or ham.
- Preprocess the raw text data and convert it into a format the machine can understand.
- Train the model using the TF-IDF technique and Logistic Regression.
- Evaluate the model's performance using standard metrics.

Ultimately, we want a reliable and fast system that could be used in real applications like email filters or messaging apps.

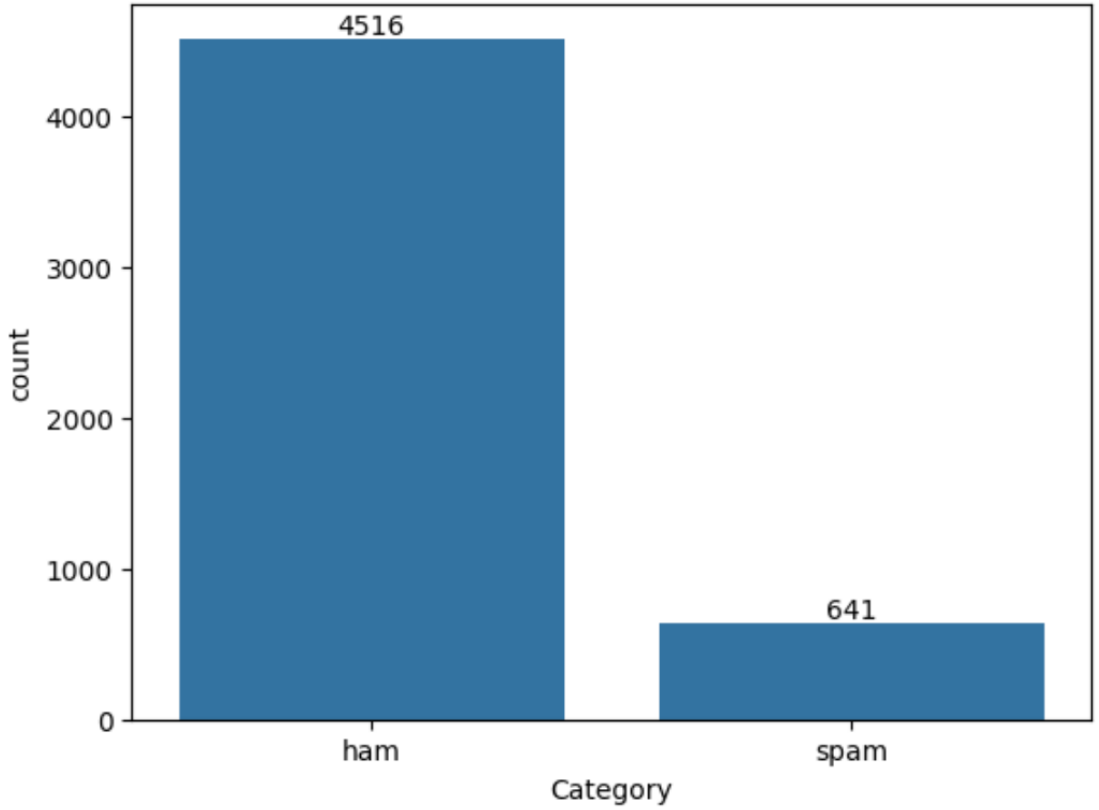
DATA OVERVIEW

The dataset contains 5,572 messages collected from SMS texts out of which 415 messages are duplicates.

It has two main columns:

- Category – the label (either spam or ham)
- Message – the actual content of the text

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...



Most messages are labelled as ham (not spam), while a smaller portion are spam, which is typical in real-world data.

INNOMATICS RESEARCH LABS

DATA PRE-PROCESSING AND MODEL BUILDING

Data Preprocessing:

- Before building the model, the text data needs to be cleaned and transformed.
- Lowercase conversion. E.g. to make sure "Free" and "free" are treated the same
- Removing special characters: Using regular expressions (regex) to remove symbols like @, #, $, !, ‘, “, ?, / etc.
- Stopword removal: Words like “the”, “is”, “a” that don’t add much meaning are removed.
- TF-IDF Vectorization: Converts each message into a numerical format by focusing on important and rare words

Model Building: The cleaned data is used to train a model

- Train-test split: 80% data used for training, 20% for testing
- SMOTE: Since spam messages were fewer than ham messages, SMOTE (Synthetic Minority Over-sampling Technique) is used to balance the classes by generating synthetic spam samples.
- Logistic Regression: Chosen because it's simple, fast, and works well for binary classification tasks like spam vs. ham.
- Grid Search CV: Used to find the best hyperparameters (like regularization strength) for the model using cross-validation to improve performance.

MODEL EVALUATION

To check how well the model performs, the following metrics are used:

- Accuracy Score: Percentage of correct predictions
- Confusion Matrix: Shows how many spam and ham messages were classified correctly or incorrectly
- Classification Report - Includes:

Precision: How many predicted spams were actually spam

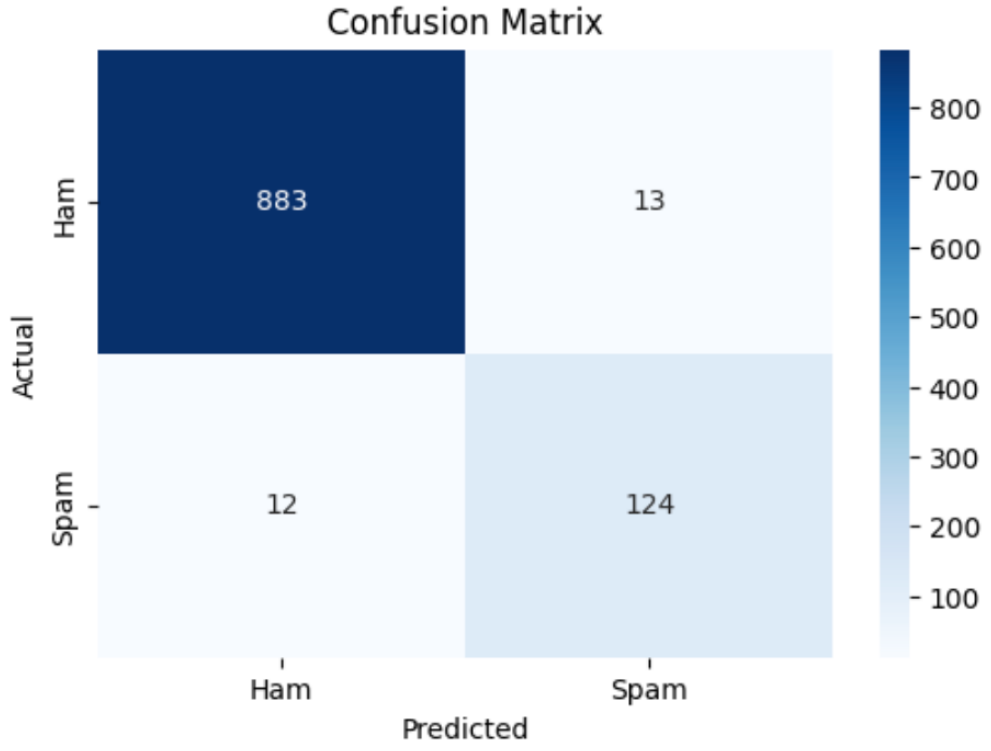
Recall: How many real spams were detected

F1-score: A balance between precision and recall

```
Model Accuracy: 0.98
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       896
           1       0.91      0.91      0.91       136

    accuracy                           0.98      1032
   macro avg       0.95      0.95      0.95      1032
weighted avg       0.98      0.98      0.98      1032
```



INNOMATICS RESEARCH LABS

Key Business Question

Can we automatically detect spam messages using machine learning so that users can avoid unwanted and potentially harmful content?

Conclusion (Key finding overall)

- The model was able to accurately detect spam messages using simple text-based features.
- TF-IDF + Logistic Regression worked well for this problem.
- Using SMOTE and GridSearchCV improved performance.
- The model achieved high precision and recall, making it reliable for real-world use.

Your Experience/Challenges Working On The Machine Learning Project

My Learning Experience:

- Learned how to clean and preprocess text data.
- Understood the importance of balancing data using SMOTE.
- Practiced model tuning with GridSearchCV.
- Learned how to evaluate performance using multiple metrics.

Challenges Faced:

- Understanding regex patterns for cleaning text
- Choosing the right vectorizer (TF-IDF vs CountVectorizer)
- Making sure the model doesn’t overfit or underperform

INNOMATICS
RESEARCH LABS

THANK YOU

INNOMATICS
RESEARCH LABS