SMS SPAM DETECTION USING MACHINE LEARNING PROJECT REPORT

SUBMITTED BY: Vihan Singh

DATA ANALYTICS COURSE

OCTOBER 2024 BATCH

INTRODUCTION

Spam has become a major issue in online communication, with around 55% of all emails being reported as spam—a number that's steadily growing. Spam, or unsolicited bulk email, allows senders to flood inboxes with unwanted ads or junk at no cost. This practice clutters millions of mailboxes worldwide, wasting time, causing users to accidentally delete legitimate emails, and even leading to economic impacts. The offensive content of spam and its disruption have prompted some countries to adopt legislation to combat it.

OVERVIEW

This project uses Logistic Regression and feature analysis to classify emails as spam or ham, helping users identify frauds, phishing attempts, and scams. By detecting spam messages with machine learning, it protects users' identity and information while preventing them from falling victim to scams.

BLOCK DIAGRAM

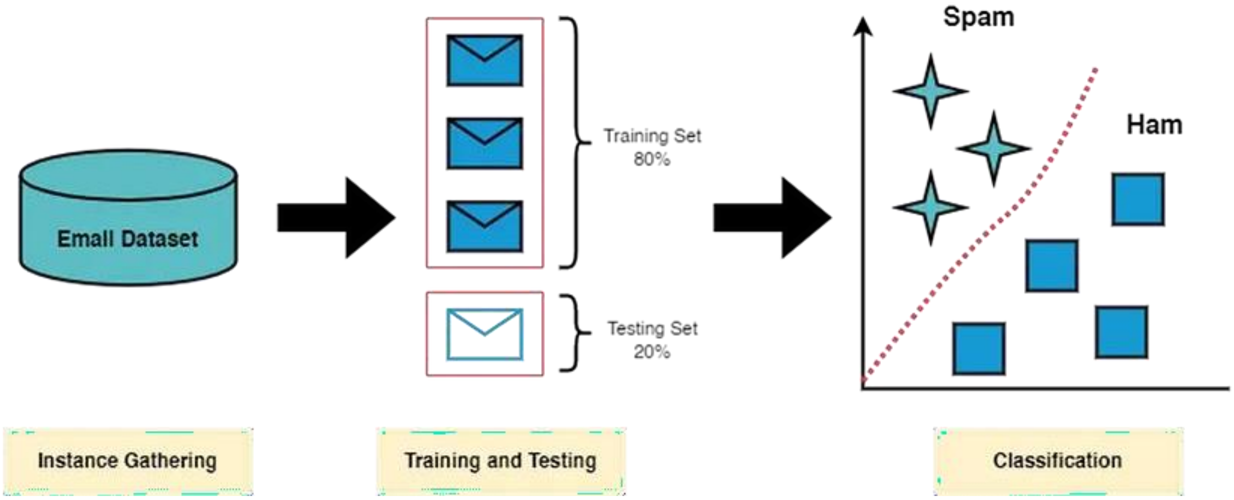


Diagram representing the flow of the system

TOOLS & LIBRARIES

- Programming Language: Python
- Data manipulation: pandas
- Machine learning: scikit-learn
- Jupyter Notebook

DATASET

The dataset, SMS Spam Collection, contains 5,574 SMS messages labeled as 'ham' or 'spam.' Each message is tagged to facilitate training and evaluation of machine learning models. Preprocessing is essential due to the presence of raw text data.

DATA PREPROCESSING

Text Cleaning: Removal of punctuation, digits, and special characters using regular expressions.

Tokenization: Splitting messages into words.

Stop Words Removal: Excluded frequently occurring words with minimal semantic value.

FEATURE ENGINEERING

Bag-of-Words (BoW): Text data transformed into numerical vectors.

TF-IDF Vectorization: Implemented for improved feature representation by accounting for term importance across the corpus.

SPLITTING DATA

The dataset was divided into training and testing subsets using an 80-20 ratio to ensure model generalization.

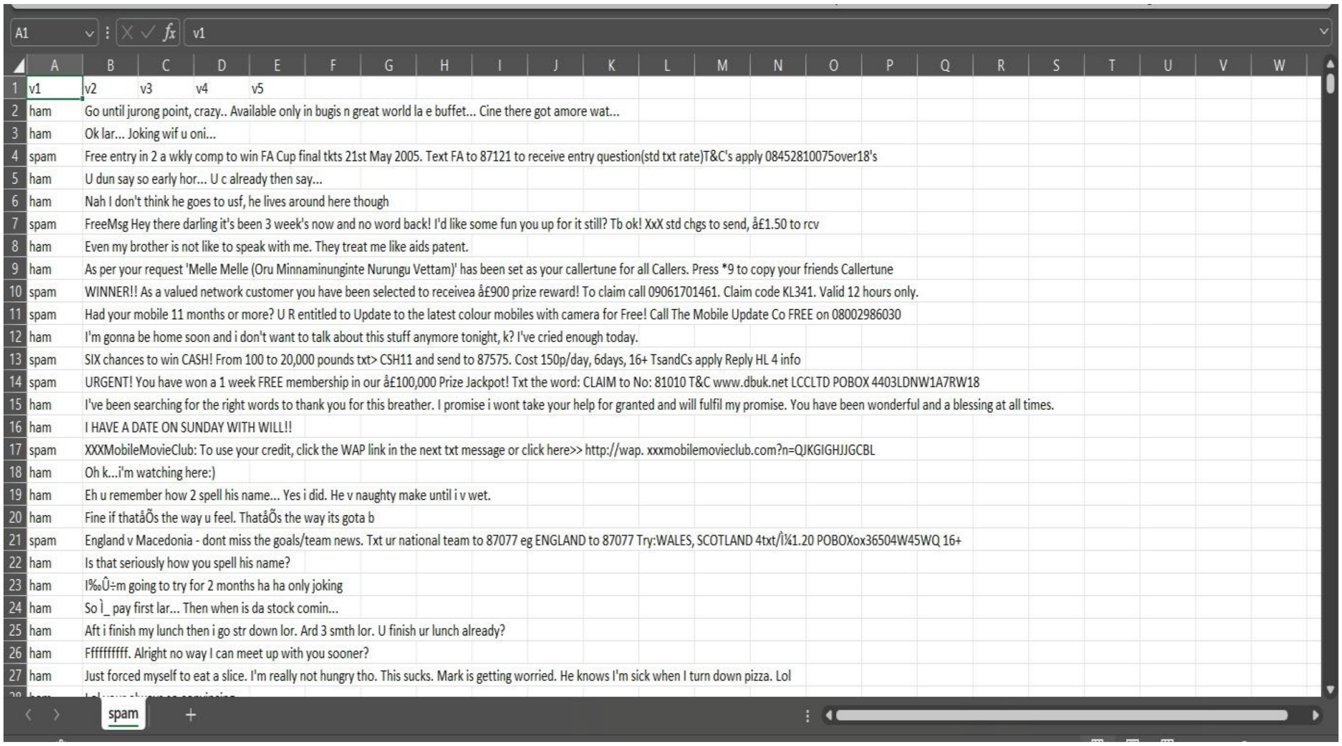
Model Development

Various classification algorithms were implemented and evaluated

TECHNIQUE USED

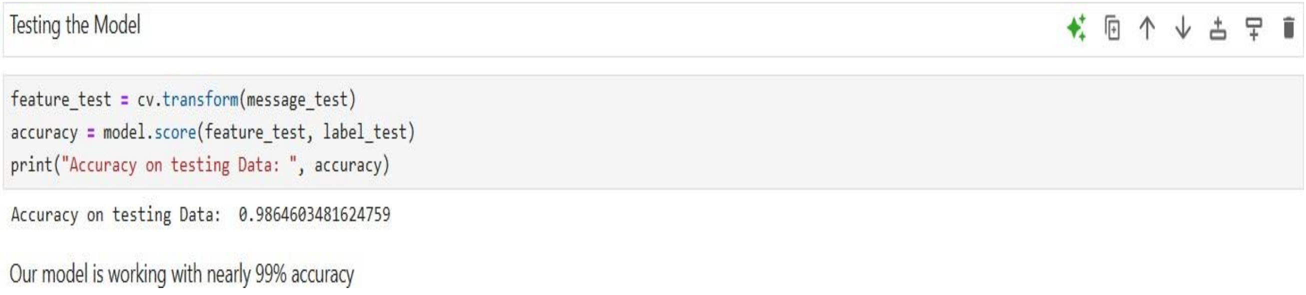
- Machine Learning
- Logistic Regression

SCREENSHOTS



	A	B	C	D	E
1	v1	v2	v3	v4	v5
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...			
3	ham	Ok lar... Joking wif u oni...			
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's			
5	ham	U dun say so early hor... U c already then say...			
6	ham	Nah I don't think he goes to usf, he lives around here though			
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv			
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.			
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune			
10	spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.			
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030			
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.			
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info			
14	spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18			
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.			
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!			
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL			
18	ham	Oh k...i'm watching here:)			
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.			
20	ham	Fine if thatåÕs the way u feel. ThatåÕs the way its gota b			
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ì¼1.20 POBOXox36504W45WQ 16+			
22	ham	Is that seriously how you spell his name?			
23	ham	I‰Û÷m going to try for 2 months ha ha only joking			
24	ham	So Ì_ pay first lar... Then when is da stock comin...			
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?			
26	ham	Ffffffffff. Alright no way I can meet up with you sooner?			
27	ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol			

Dataset Image



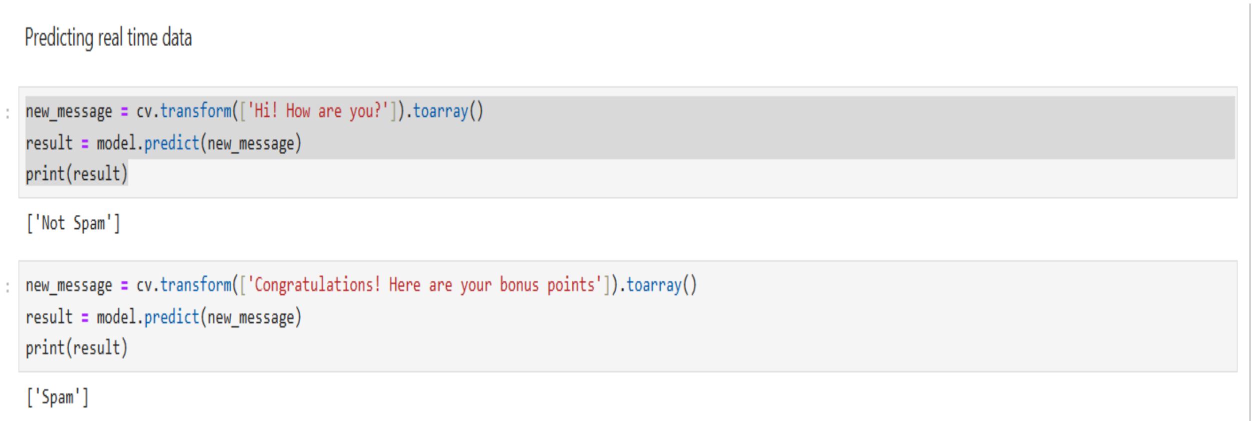
Testing the Model

```
feature_test = cv.transform(message_test)
accuracy = model.score(feature_test, label_test)
print("Accuracy on testing Data: ", accuracy)
```

```
Accuracy on testing Data:  0.9864603481624759
```

Our model is working with nearly 99% accuracy

Accuracy Score



Predicting real time data

```
new_message = cv.transform(['Hi! How are you?']).toarray()
result = model.predict(new_message)
print(result)
```

```
['Not Spam']
```

```
new_message = cv.transform(['Congratulations! Here are your bonus points']).toarray()
result = model.predict(new_message)
print(result)
```

```
['Spam']
```

Prediction Result

TESTING

Testing is the process of evaluation of a system to detect differences between given input and expected output and also to assess the feature of the system. Testing assesses the quality of the product. It is a process that is done during the development process.

CONCLUSION

The SMS Spam Detection project effectively demonstrates how Natural Language Processing (NLP) and machine learning techniques can be applied to solve real-world problems like spam filtering. By leveraging a comprehensive dataset and systematically implementing preprocessing, feature engineering, and model evaluation techniques, this project achieved accurate classification of SMS messages into spam and ham categories.

The project underscores the value of rigorous preprocessing and systematic evaluation in developing high-performing NLP models. With further enhancements, such as using deep learning techniques or expanding to multi-language datasets, this model can be adapted for broader spam detection applications, paving the way for more robust and scalable solutions.