PROJECT PRESENTATION

CSE445.7

E-MAIL / SMS SPAM CLASSIFICATION

An end-to-end website through ML to detect spam and not spam messages.

PRESENTED BY

Rahul Deb Roy (1931132042) &
Moriom Islam Mou (1931333042)

AGENDA

1. Objectives
2. Description Of The Dataset
3. Data Cleaning & Preprocessing
4. Model Building
5. Model Evalutions
6. Improvements
7. Results
8. Motivation

OBJECTIVES

- To develop algorithms or models that can automatically identify and classify incoming emails as either spam or legitimate (non-spam).

- To provide users with a more efficient and secure email experience by reducing the volume of spam and ensuring that important messages are not mistakenly identified as spam thus reducing the amount of unwanted or malicious content that users have to deal with.

DESCRIPTION OF DATASET

Source: **Kaggle**

	Type	Text
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...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

5572 rows × 2 columns

Domain	Public
Attributes	2 - (Type & Text)
NO. of Rows	5572
Size	5572 observations and 2 features
Target Attribute	Type
Target Class	Ham & Spam

DATA CLEANING & PREPROCESSING

Number of Missing/ Null values

0

```
# Checking for missing/null values
df.isnull().sum()
```

```
Type    0
Text    0
```

Labelling Target Classes

Ham = 0 & **Spam** = 1

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

Duplicate values

415

Size of Updated DS

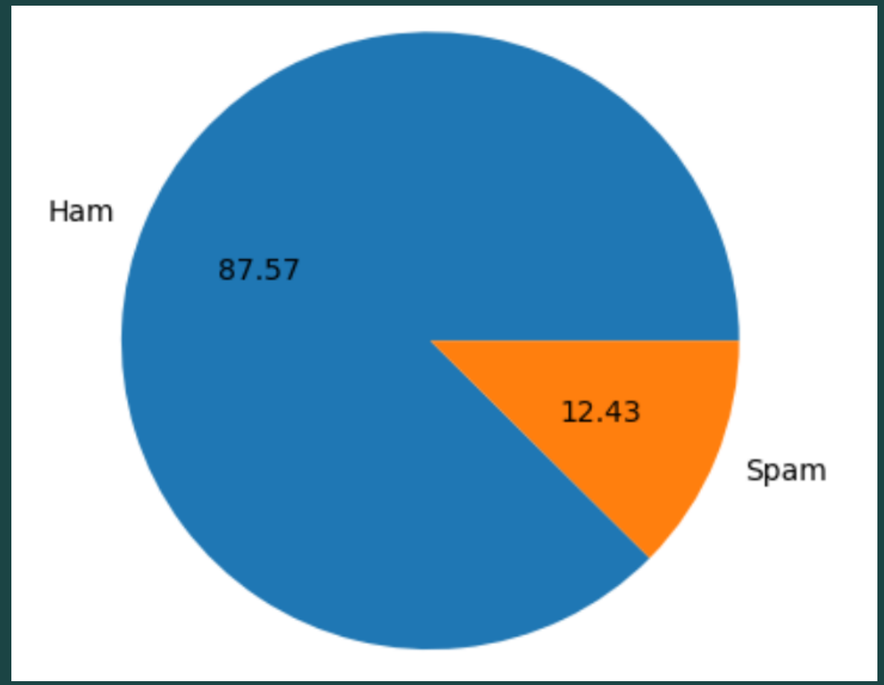
5157 rows × 2 columns

Number of Ham texts

4516

Number of Spam texts

641



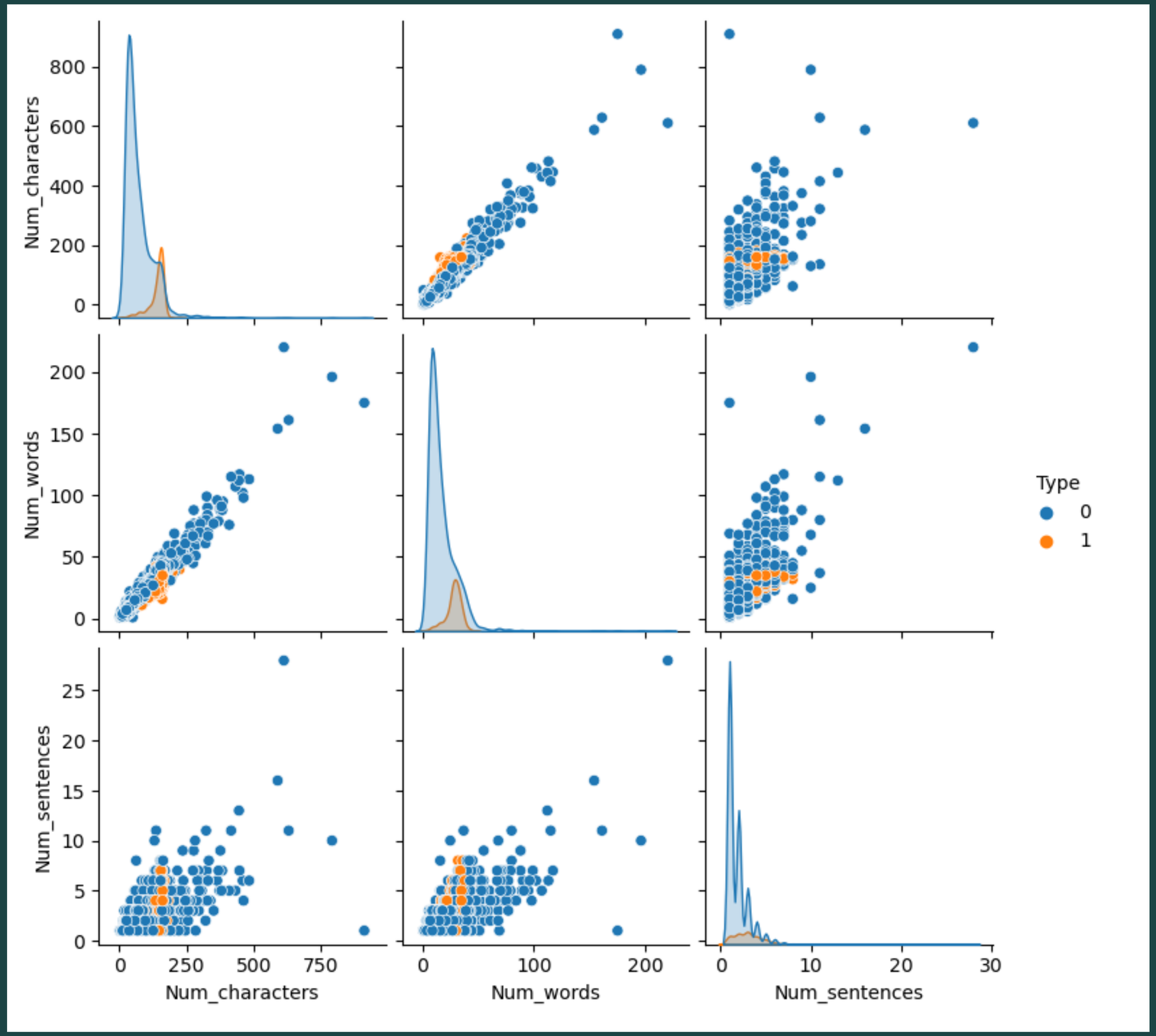
	Type	Text	Num_characters	Num_words	Num_sentences
0	0	Go until jurong point, crazy.. Available only ...	111	24	2
1	0	Ok lar... Joking wif u oni...	29	8	2
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2
3	0	U dun say so early hor... U c already then say...	49	13	1
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1

For Ham Messages

	Num_characters	Num_words	Num_sentences
count	4516.000000	4516.000000	4516.000000
mean	70.869353	17.264836	1.806244
std	56.708301	13.587852	1.281910
min	2.000000	1.000000	1.000000
25%	34.000000	8.000000	1.000000
50%	53.000000	13.000000	1.000000
75%	91.000000	22.000000	2.000000
max	910.000000	220.000000	28.000000

For Spam Messages

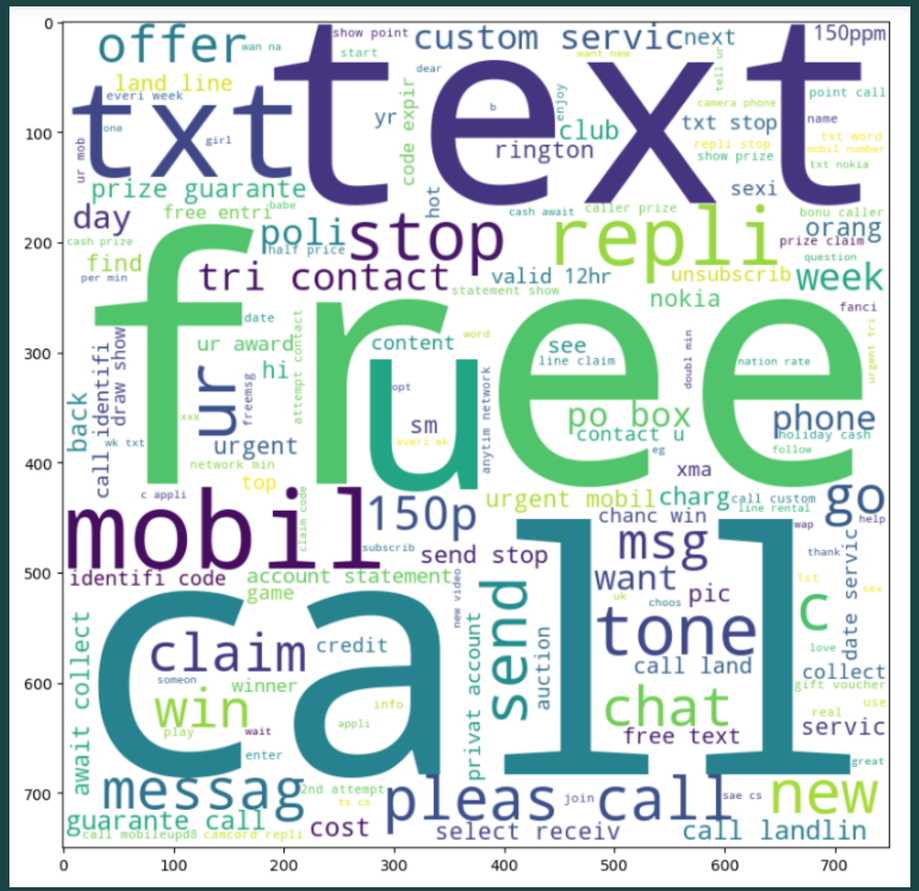
	Num_characters	Num_words	Num_sentences
count	641.000000	641.000000	641.000000
mean	137.118565	27.667707	2.967239
std	30.399707	7.103501	1.480241
min	7.000000	2.000000	1.000000
25%	130.000000	25.000000	2.000000
50%	148.000000	29.000000	3.000000
75%	157.000000	32.000000	4.000000
max	223.000000	46.000000	8.000000


Num_characters
Num_words
Num_sentences
Type
0
1
0
250
500
750
0
100
200
0
10
20
30
0
5
10
15
20
25
0
50
100
150
200
0
200
400
600
800

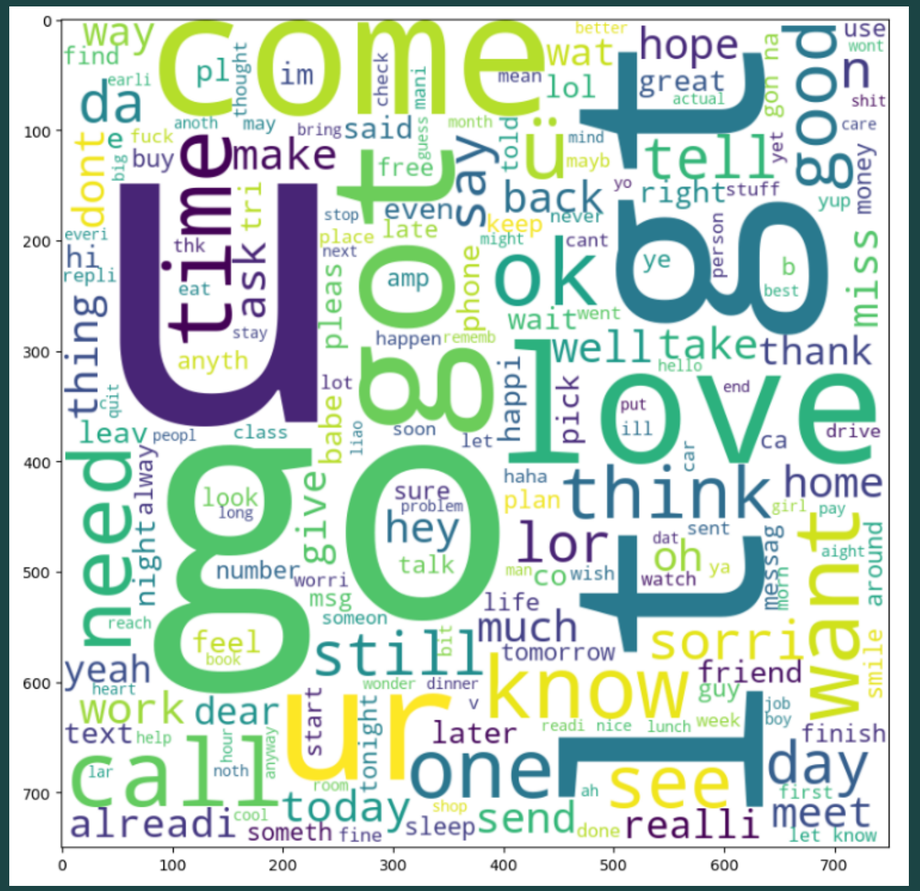
	Type	Text	Num_characters	Num_words	Num_sentences	Transformed_text
0	0	Go until jurong point, crazy.. Available only ...	111	24	2	go jurong point crazi avail bugi n great world...
1	0	Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
3	0	U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

5157 rows × 6 columns

For Spam Messages



For Ham Messages



MODEL BUILDING

GaussianNB	MultinomialNB
BernoulliNB	LogisticRegression
SVM	Decision Tree Classifier
K Neighbors Classifier	RandomForestClassifier
ExtraTreesClassifier	Voting Classifier

MODEL EVALUATIONS

	Accuracy	Precision	Recall	Confusion Matrix
GaussianNB	• 0.86	• 0.47	• 0.86	• [[784, 121] [17, 110]]
MultinomialNB	• 0.97	• 1.0	• 0.76	• [[905, 0] [30, 97]]
BernoulliNB	• 0.98	• 0.98	• 0.88	• [[903, 2] [15, 112]]
SVM	• 0.97	• 0.98	• 0.81	• [[903, 2] [23, 104]]
K Neighbors Classifier	• 0.91	• 1.0	• 0.29	• [[905, 0] [90, 37]]

MODEL EVALUATIONS

	Accuracy	Precision	Recall	Confusion Matrix
Decision Tree Classifier	• 0.93	• 0.82	• 0.61	• [[888 17] [49 78]]
LogisticRegression	• 0.95	• 0.93	• 0.70	• [[899 6] [37 90]]
RandomForestClassifier	• 0.97	• 1.0	• 0.78	• [[905 0] [27 100]]
ExtraTreesClassifier	• 0.97	• 0.99	• 0.82	• [[904 1] [22 105]]

MODEL IMPROVEMENT

Voting Classifier

Accuracy	**0.98**
Precision	**1.0**
Recall	**0.86**
Confusion matrix	**[[905 0] [18 109]]**

RESULTS


Email/SMS Spam Classifier
Enter the message
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, £1.50 to rcv
Predict
Spam


Email/SMS Spam Classifier
Enter the message
Hi, can you come over for dinner?
Predict
Not Spam

MOTIVATION

By classifying emails as spam or non-spam, email providers and users can effectively manage their email communications, prioritize important messages, and protect themselves from potential phishing attempts, scams, or malware that may be contained within spam emails.

Thank you!