

#### SPAMMAIL/SMS CLASSIFICATION

Presented by Team Threes

# Today's Agenda

- Team Introduction
- Problem Statement
- Impact of the problem
- Objective
- Project Scope
- Demo
- Al ethics
- Future Work

#### The Team

Thwet Yin Nyein

Yu Shwe Yee Htet

Paing Pyih Ko

### What is Spam?



Any kind of unwanted, unsolicited digital communication that gets sent out in bulk.

Often spam is sent via email, but it can also be distributed via text messages, phone calls, or social media.

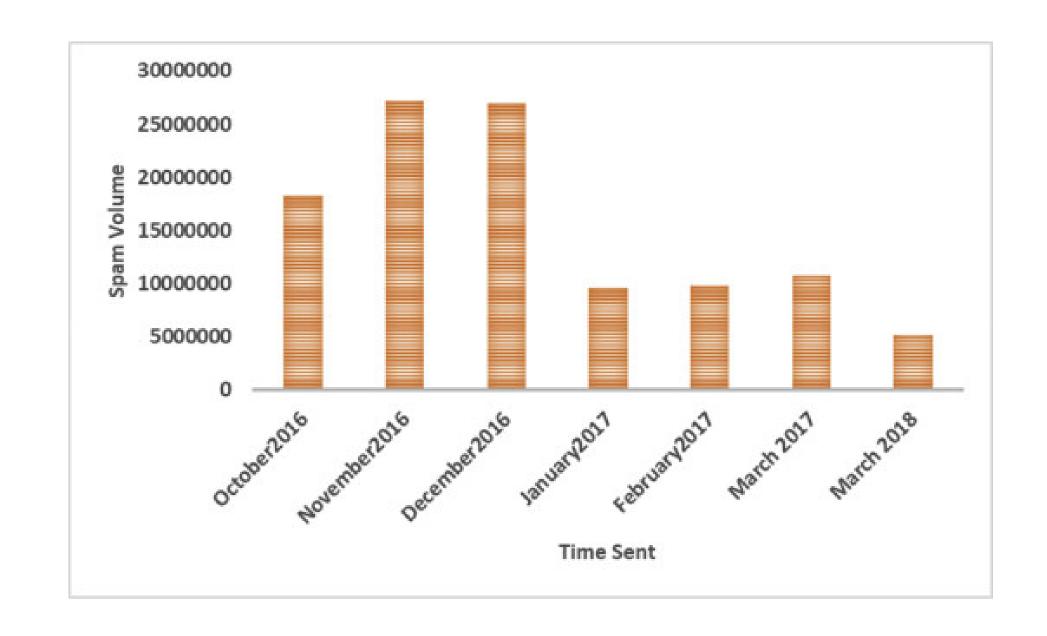
#### What is Spam?

- Nearly 85% of all emails are spam.
- Costs businesses \$20.5 billion every year
- Microsoft accounts are the most popular targets of phishing emails, accounting for 43% of all phishing attempts.
- Americans admit to losing more than \$70,000 to Nigerian Prince scams in 2019.

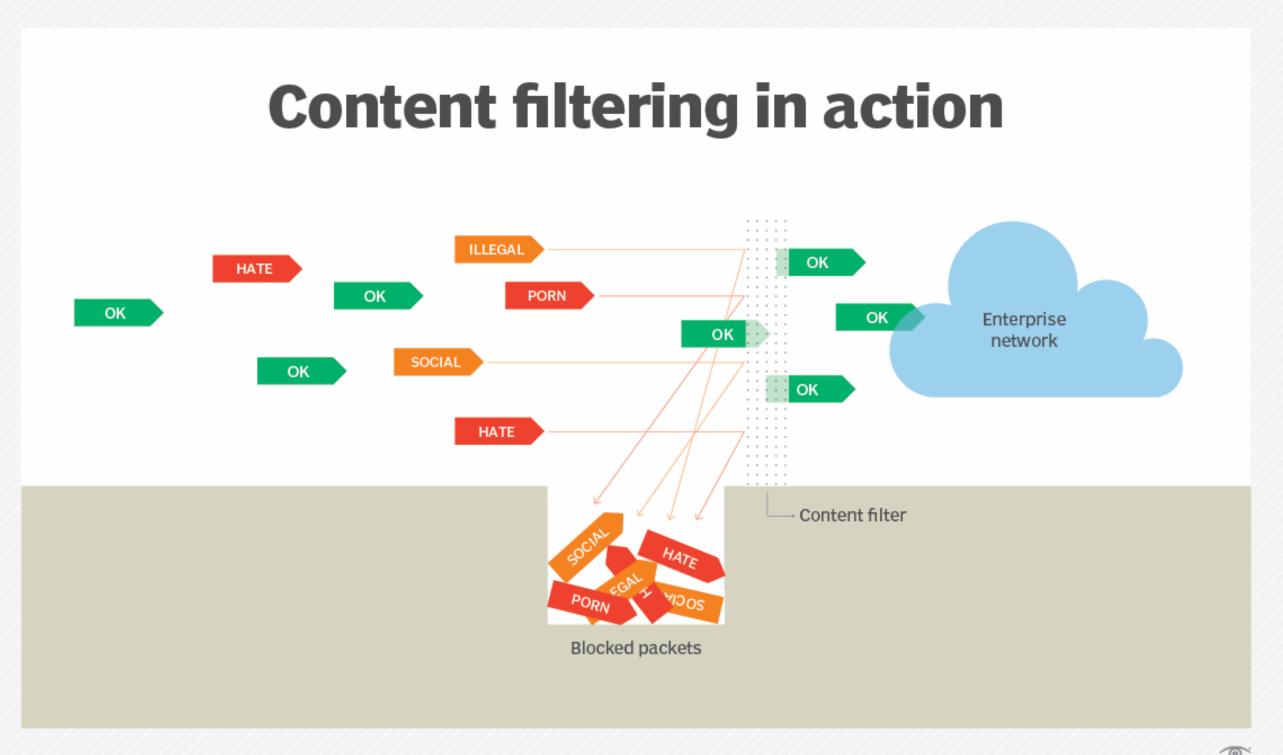
# Spam Volume across the years

Destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time .

The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic.



# Project Purpose

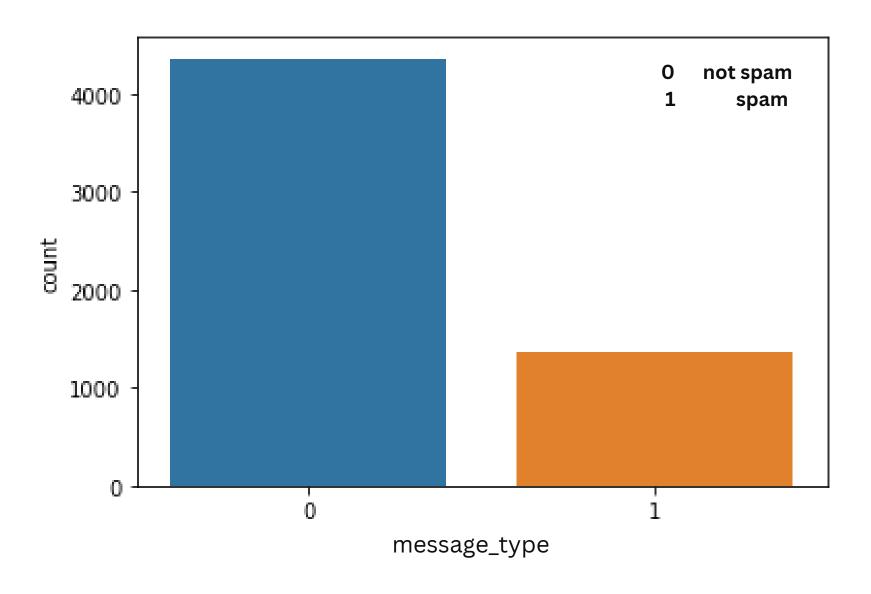


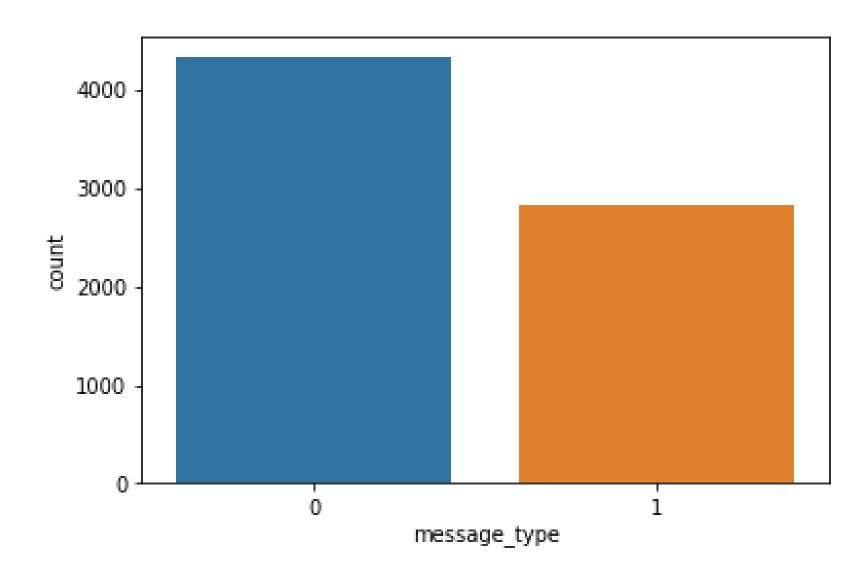
# Project Timeline



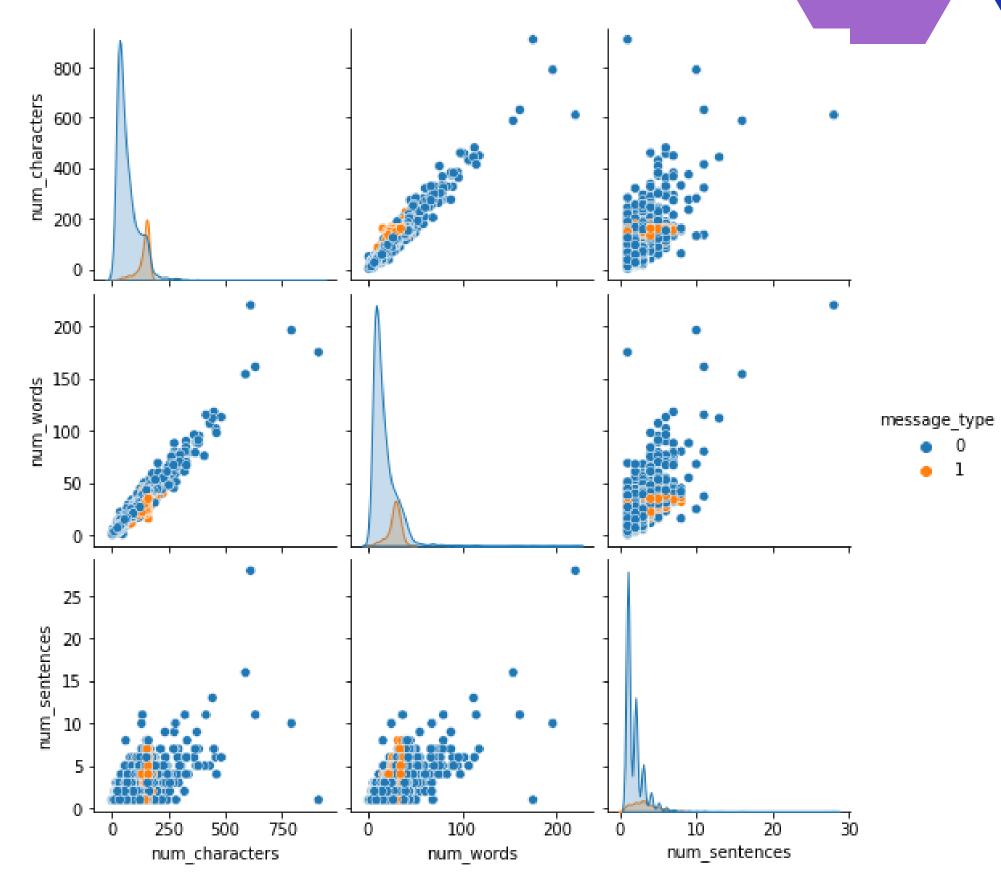
#### Data Collection

• using CSV files



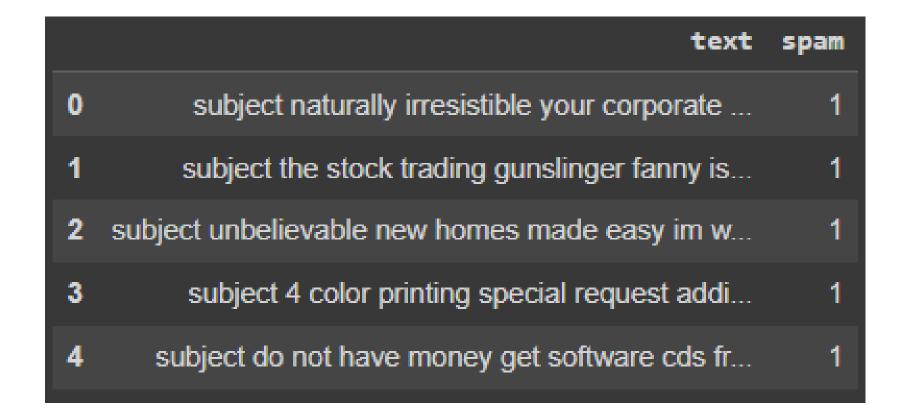


#### **Exploratory Data Analysis**



#### Tokenization

```
[Subject, :, naturally, irresistible, your, co...]
[Subject, :, the, stock, trading, gunslinger, ...]
[Subject, :, unbelievable, new, homes, made, e...]
[Subject, :, 4, color, printing, special, requ...]
[Subject, :, do, not, have, money, ,, get, sof...]
```



```
Go jurong point crazy Available bugis n great ...
                                 Ok lar Joking wif u oni
                     U dun say early hor U c already say
             Nah dont think goes usf lives around though
           Even brother like speak treat like aids patent
       Wealth without risk discover best kept secret ...
6954
       Sir madan strictly confidential pleased introd...
6955
       mr desmond stevens urgent assistance may surpr...
6956
       Abc good morning america ranks NUMBER christma...
6957
       hyperlink hyperlink let mortgage len...
6958
Name: message, Length: 6410, dtype: object
```

Remove stop words & punctuations

	message_type	message	num_characters	num_words	num_sentences	count
0	0	go jurong point crazy available bugis n great	111	24	2	24
1	0	ok lar joking wif u oni	29	8	2	8
2	0	u dun say early hor u c already say	49	13	1	13
3	0	nah dont think goes usf lives around though	61	15	1	15
4	0	even brother like speak treat like aids patent	77	18	2	18

Stemming

a = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

The	quick	brown	fox	jumps	over	lazy	dog
2	1	1	1	1	1	1	1

Count Vectorizer-transform text to numerical data

#### **Bag of Words**

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

• simply counts the frequency of words in a document

#### TF-IDF

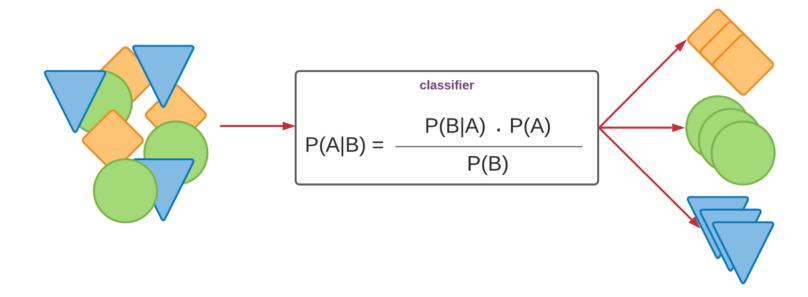
Document 1: Text processing is necessary.

Document 2: Text processing is necessary and important.

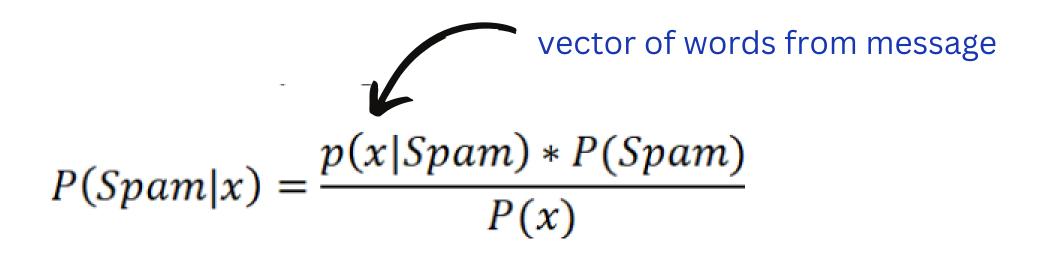
Word	TF		IDF	TFIDF		
	Doc 1	Doc 2		Doc 1	Doc 2	
Text	1/4	1/6	log(2/2) = 0	0	0	
Processing	1/4	1/6	log (2/2) =0	0	0	
ls	1/4	1/6	log (2/2) =0	0	0	
Necessary	1/4	1/6	log (2/2) =0	0	0	
And	0/4	1/6	log (2/1) =0.3	0	0.05	
Important	0/4	1/6	log (2/1) =0.3	0	0.05	

# Model-Training

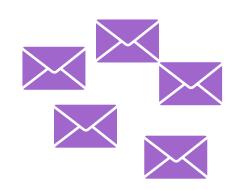
### Multinomial Naive Bayes



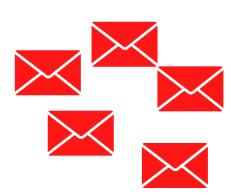
- The multinomial model captures word frequency information in documents
- Capturing frequency information of tokenization can help classification
- Require a small number of training sets but effective classification results
- The more applicable algorithm in text classification
- The most popular among naïve bayes classifiers



x=[w1,w2,w3,w4,...,wn]

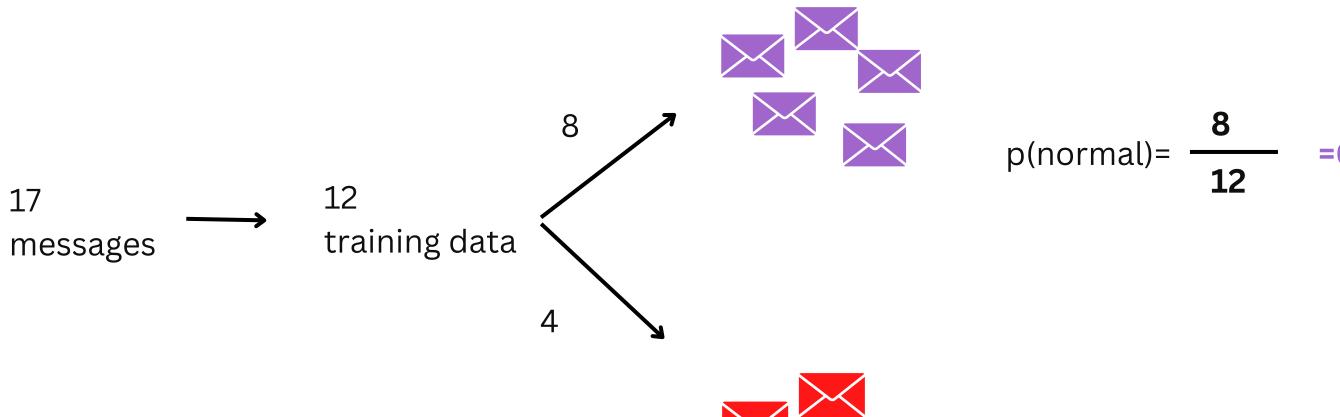


Word	Count	
Dear	9	p(Dear  normal) = 9/17 = 0.52
Friend	6	p(Friend normal) = 6/17 = 0.35
Promotion	1	p(promotion  normal)= 1/17 = 0.29
Free	1	p(free  normal) = $1/17 = 0.29$
Total	<b>17</b>	

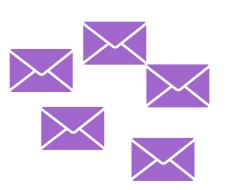


Word	Count	
Dear	1	p(Dear  spam) = $1/7 = 0.14$
Friend	1	p(Friend spam) = 1/7 = 0.14
Promotion	1	p(promotion  spam)= 1/7 = 0.14
Free	4	p(free  spam) = $4/7 = 0.57$
Total	7	

likelihoods



p(spam)= 
$$\frac{4}{12}$$
 =0.33



```
p(Dear| normal) = 8/17 = 0.47

p(Friend| normal) = 5/17 = 0.29

p(promotion| normal) = 3/17 = 0.18

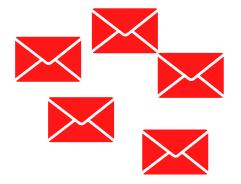
p(free| normal) = 1/17 = 0.29
```

Dear Friend

```
p(normal)=0.67 p(r)
```

p(normal) \* p(Dear|normal) \* p(Friend|normal)





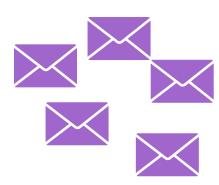
```
p(Dear| spam) = 1/7 = 0.14
p(Friend| spam) = 1/7 = 0.14
p(promotion| spam)= 1/7 = 0.14
p(free| spam) = 4/7 = 0.57
```

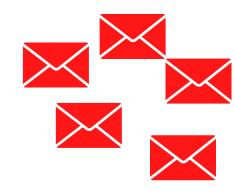
p(spam)=0.33 p(spam) \* p(Dear|spam) \* p(Friend|spam)



0.33 \* 0.14 \* 0.14= 0.006







```
p(Dear|normal) = 8/17 = 0.47

p(Friend|normal) = 5/17 = 0.29

p(promotion|normal) = 3/17 = 0.18

p(free|normal) = 1/17 = 0.29
```

**Promotion Free Free** 

```
p(Dear| spam) = 1/7 = 0.14
p(Friend| spam) = 1/7 = 0.14
p(promotion| spam)= 1/7 = 0.14
p(free| spam) = 4/7 = 0.57
```

```
p(normal) * p(promotion|normal) * p(free|normal) * p(free|normal)
```

p(spam) \* p(promotion|spam) \* p(free|spam) \* p(free|spam)

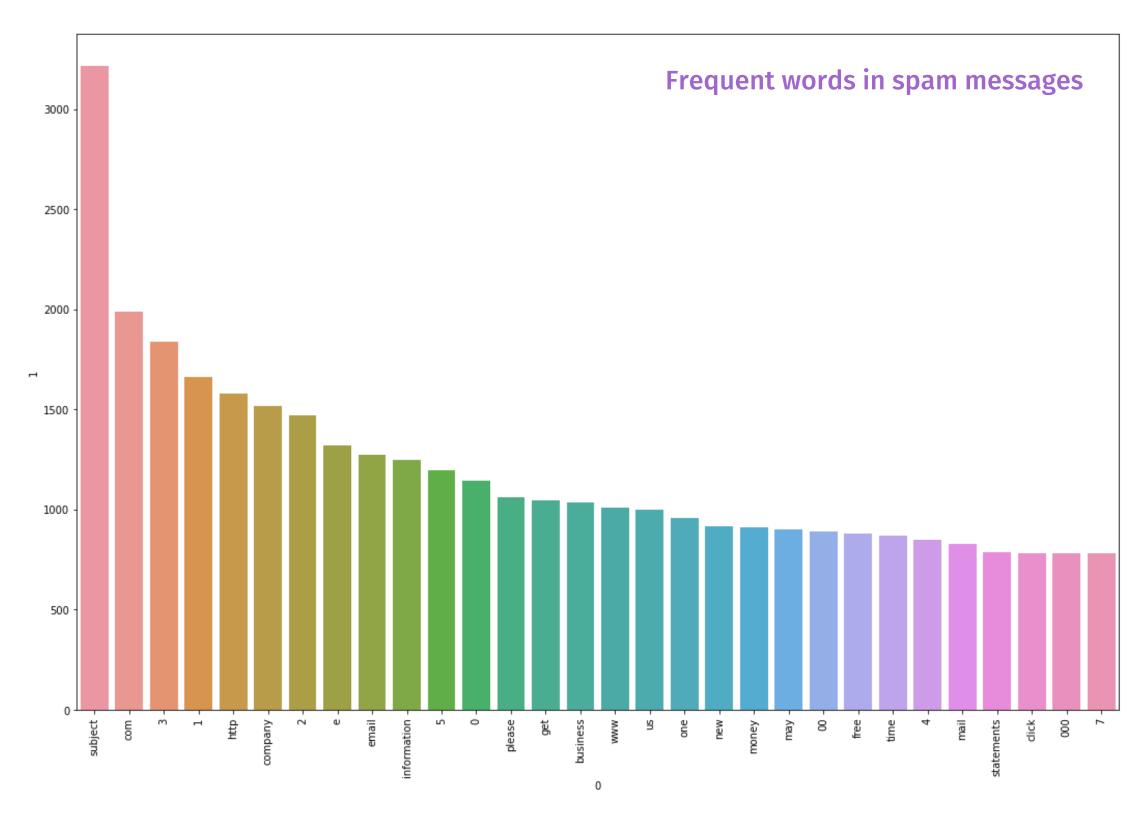
0.18\* 0.29\*0.29= 0.015



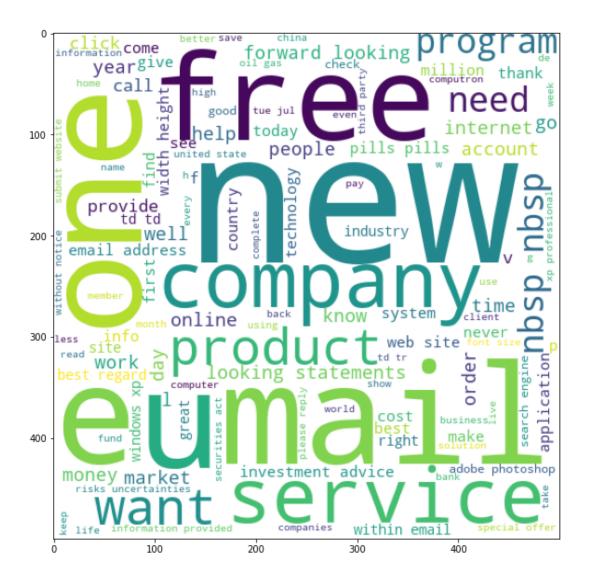
0.14\*0.57\*0.57= 0.045

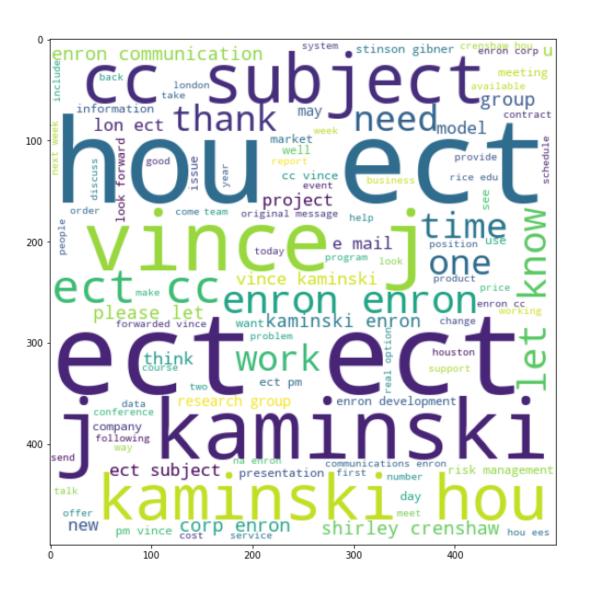
Promotion Free Free

#### Data visualization



#### Data visualization

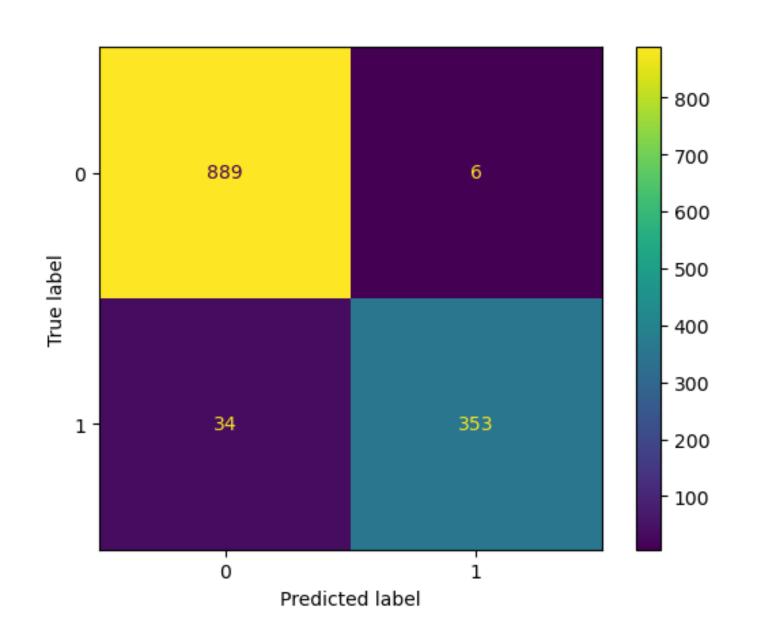




spam

not spam

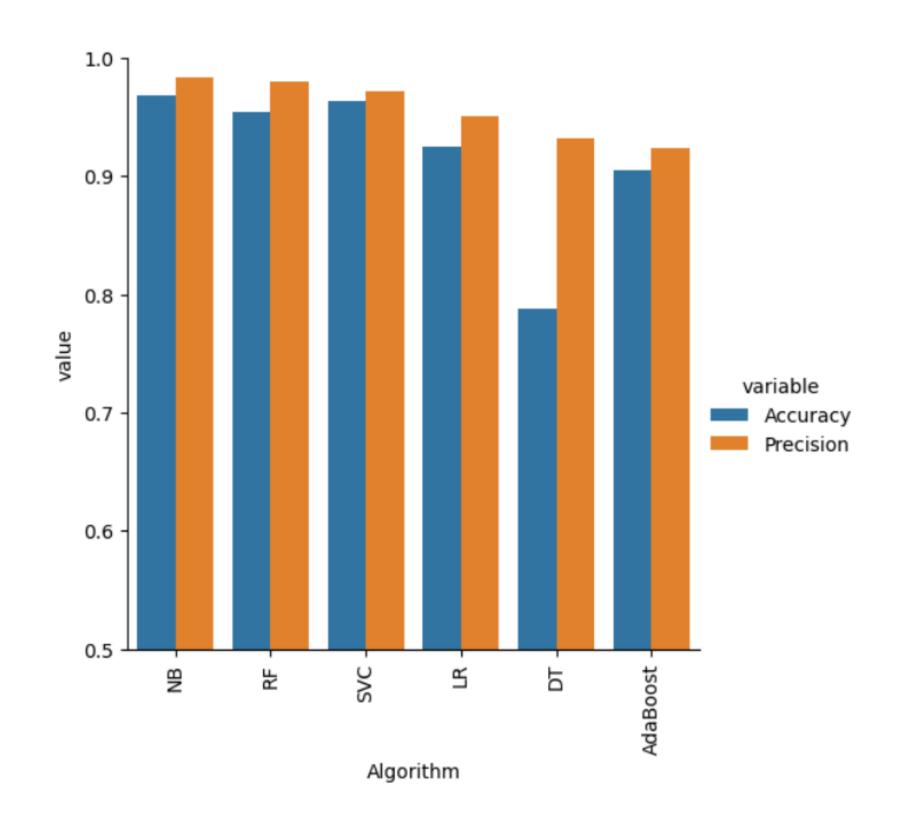
### Experimental Result



```
Accuracy score of Multinomial NB is: 0.968798751950078
Confusion Matrix of Multinomial NB is: [[889 6]
[34 353]]
Precision score of the Multinomial NB is 0.9832869080779945
```

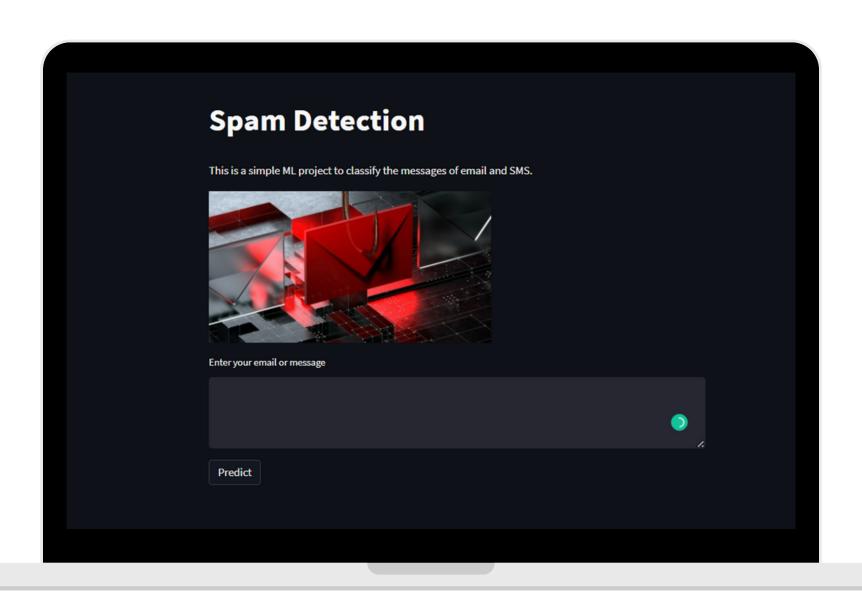
**Confusion matrix** 

#### Model Evaluation



	Algorithm	Accuracy	Precision
1	NB	0.968799	0.983287
4	RF	0.953978	0.979532
0	SVC	0.964119	0.972299
3	LR	0.925117	0.950464
2	DT	0.787832	0.932331
5	AdaBoost	0.905616	0.923567

# Streamlit app



#### Al Ethics

#### Fairness - To avoid learning Bias

These major steps are used in the functioning.

- morphological and Lexical Analysis
- independent to racists words

#### For model improvement

- Ambiguity- the phrases used have more than one meaning
- Errors in text and speech- to encounter a lack of efficiency when expose to misspelled words
- Usage of Slang and Colloquial words- high chance
- Semantic Analysis

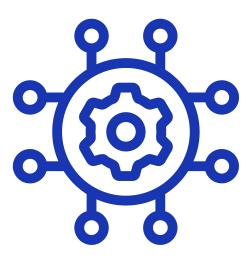
#### Future Work



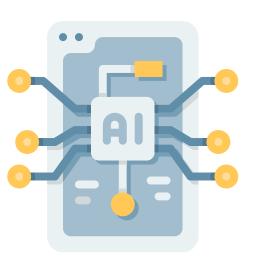
**Enrich Data** 



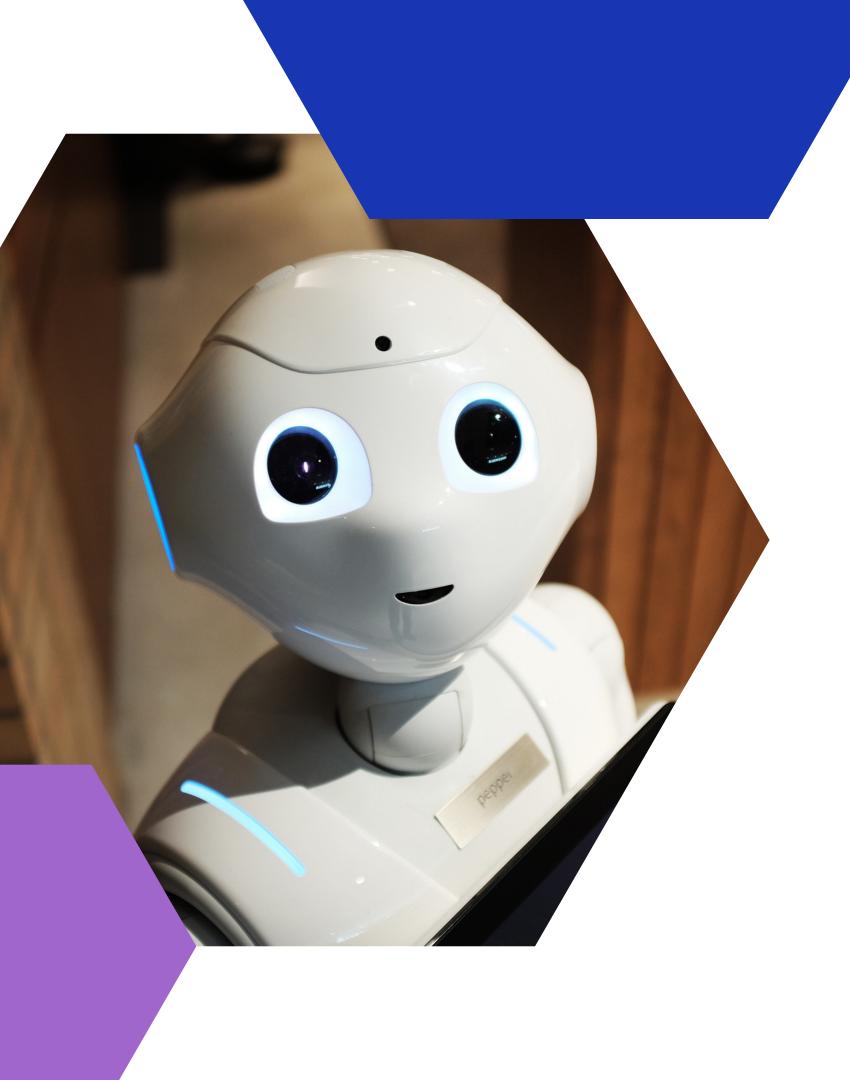
Preprocessing techniques



Improve model



sentiment analysis app



#### Thank You!

And kudos for all your hard work.