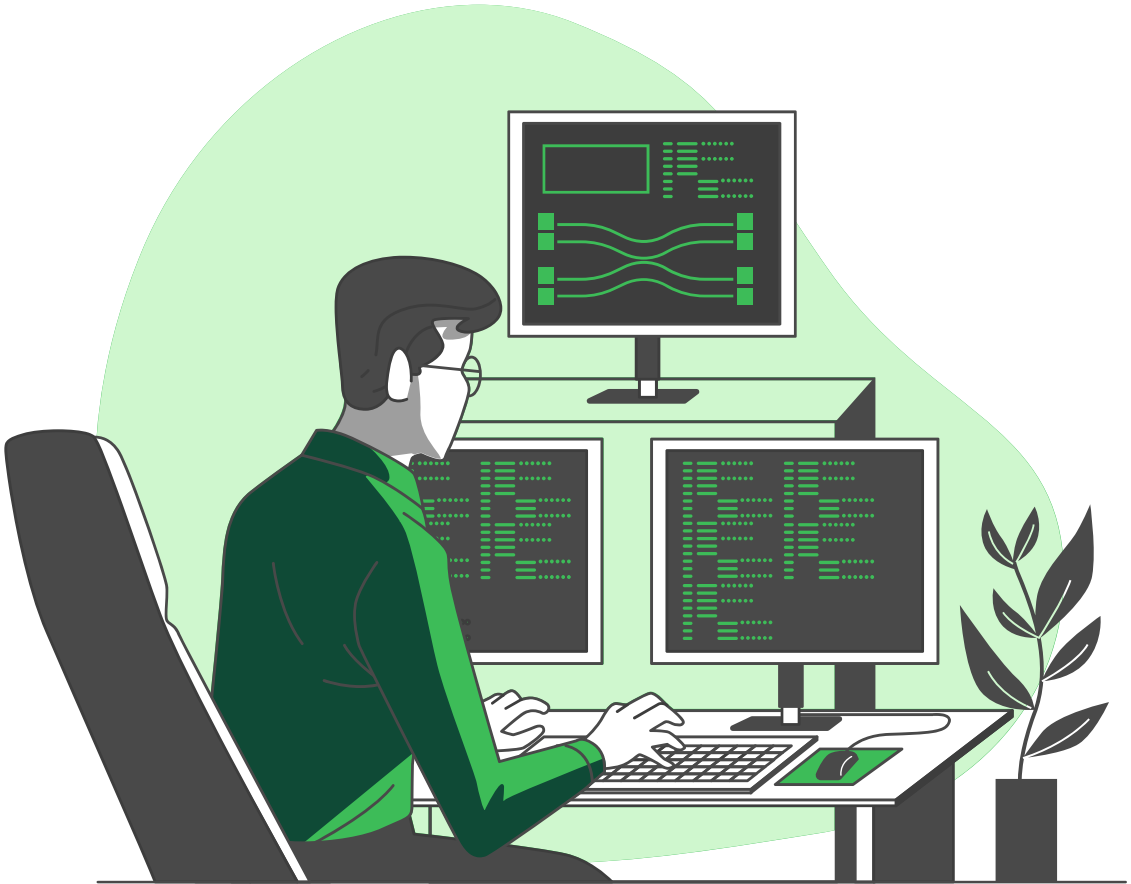
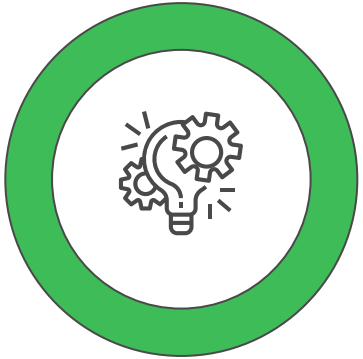


CSC447 Final Project

Emily Kim



Email Spam Classification

Focus on classifying emails as spam or non-spam

Dataset

Kaggle

Dataset 1
Dataset 2

Dataset Details

Dataset 1:
2 columns, 5k rows
Columns: Category, Message

Dataset 2:
2 columns, 3k rows
Columns: Email, Label

Combining

Final Dataset:
2 columns, 8k rows
Columns: Label, Email

Final Dataset

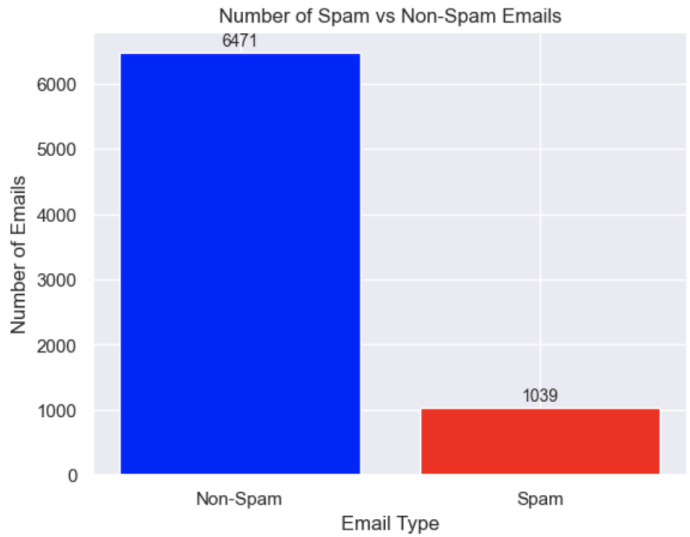
	label	email
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

0 = non-spam/ham
1 = spam

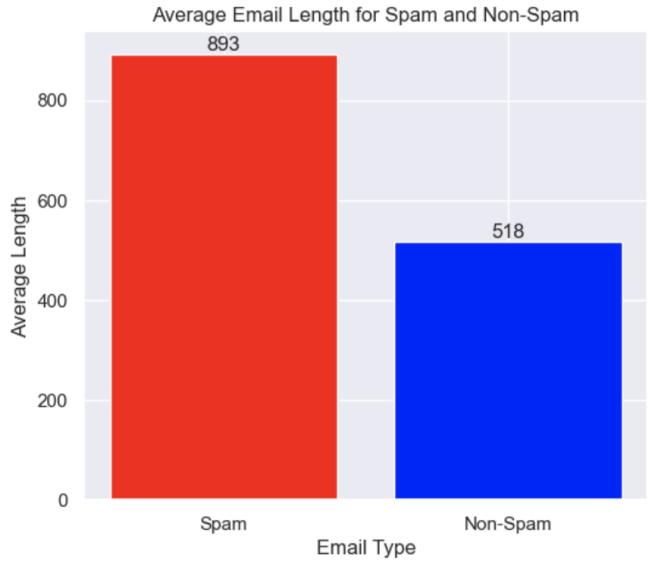
Email message

Dataset Analysis

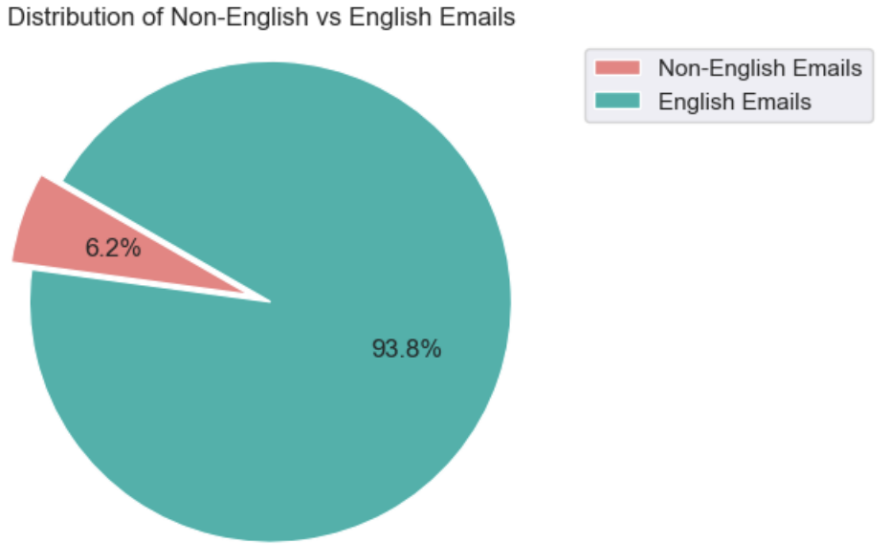
Number of Emails



Average Length



Handling Non-English Text



- There were 496 non-English in this dataset. 467 of them being non-spam and 29 being spam.

Preprocessing

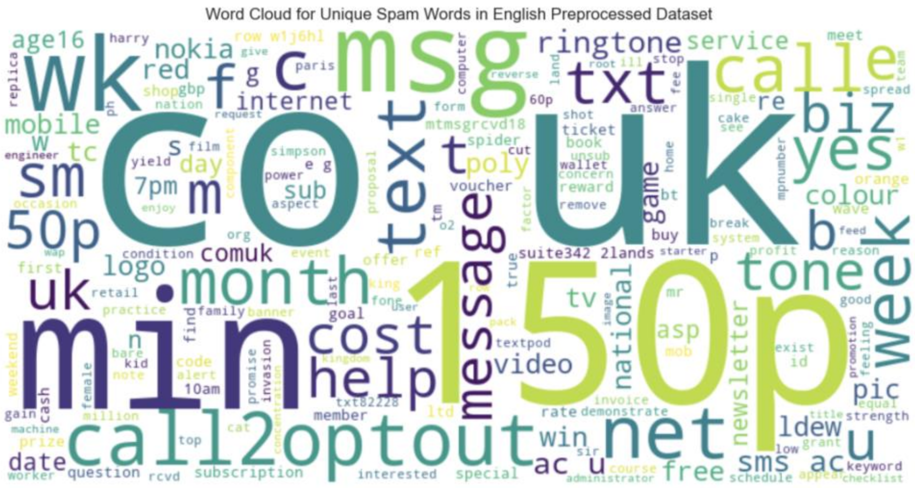
[91]:

	label	email	processed_email
0	0	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...	[go, jurong, point, crazy, .., available, bugis, n, great, world, la, e, buffet, ..., cine, got, amore, wat, ...]
2	1	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	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005., text, fa, 87121, receive, entry, question, std, txt, rate, c, 's, apply, 08452810075over18, 's]
3	0	U dun say so early hor... U c already then say...	[u, dun, say, early, hor, ..., u, c, already, say, ...]
4	0	Nah I don't think he goes to usf, he lives around here though	[nah, n't, think, goes, usf, lives, around, though]
5	1	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	[freemsg, hey, darling, 's, 3, week, 's, word, back, 'd, like, fun, still, tb, ok, xxx, std, chgs, send, £1.50, rcv]

- Lowered the text to ensure uniformity
- Broke down the text into individual words
- Removed stop words and punctuations

Word Cloud Analysis

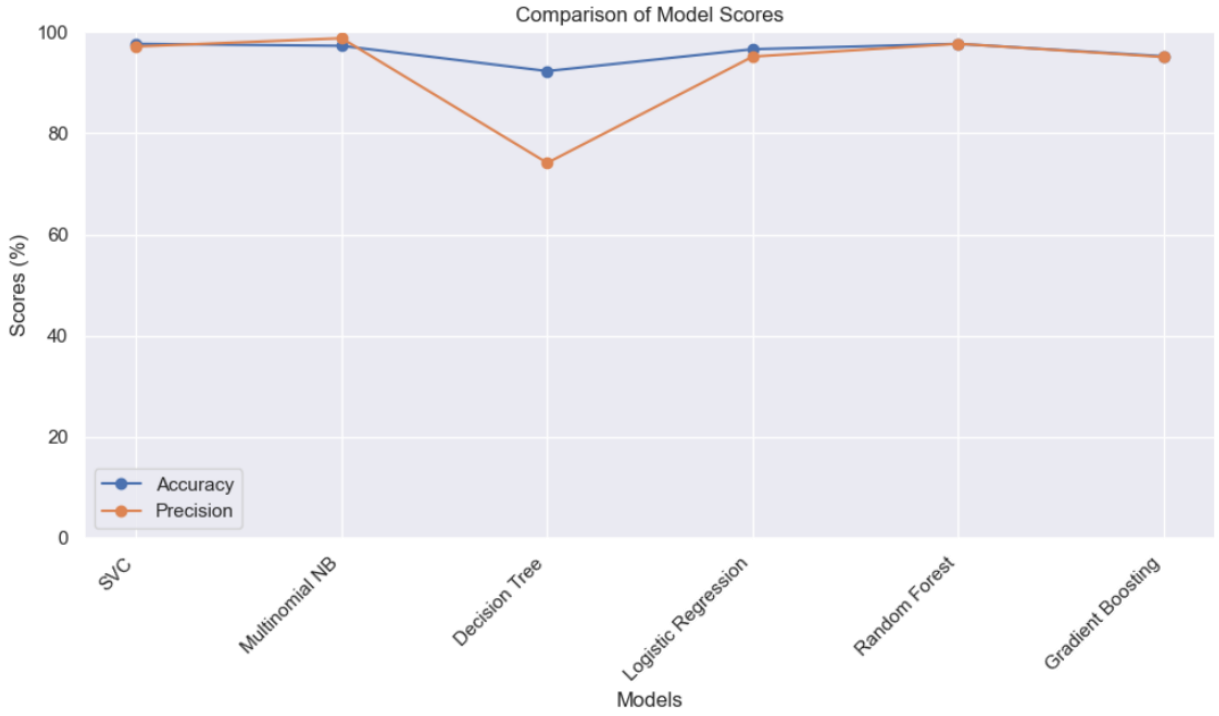
Spam Words



Non-Spam Words

Models

Model	Accuracy	Precision
SVC	0.98	0.97
Multinomial NB	0.97	0.99
Decision Tree	0.92	0.74
Logistic Regression	0.97	0.95
Random Forest	0.98	0.98
Gradient Boosting	0.95	0.95



Models

SVC

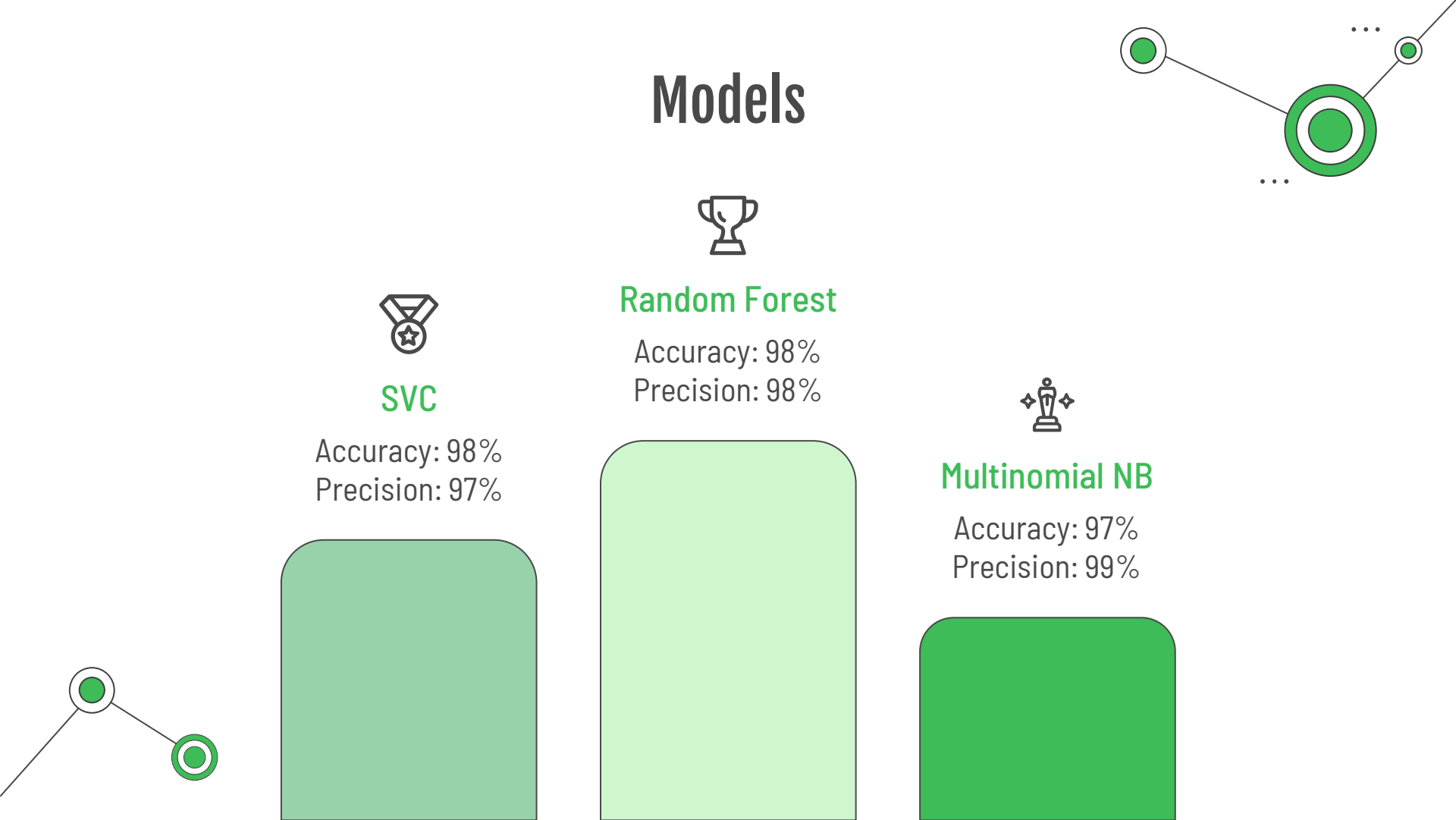
Accuracy: 98%
Precision: 97%

Random Forest

Accuracy: 98%
Precision: 98%

Multinomial NB

Accuracy: 97%
Precision: 99%



Making Predictions

Example 1: Congratulations! Click here to get your reward!
Category: This is a SPAM email.

Example 2: When do you want to meet up?
Category: This is NOT a spam email.

Example 3: new page NUMBER hyperlink hyperlink finally a newsfeed that delivers current and relevant sales marketing advertising articles from such magazines as business...
Category: This is a SPAM email.

Example 4: bees don't care what humans think is impossible
Category: This is NOT a spam email.

Thanks!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Stories