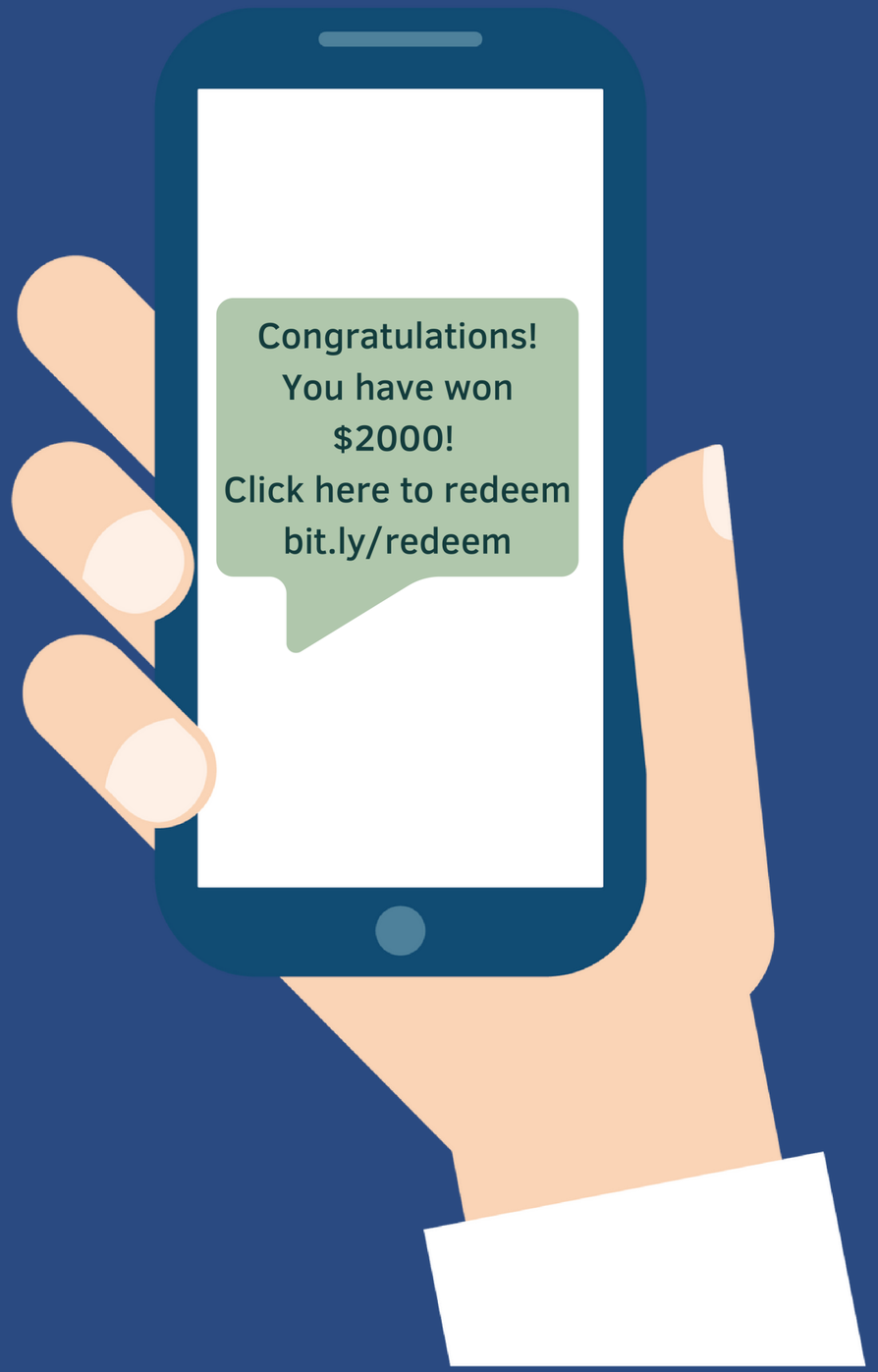
SC1015 PROJECT

SPAM BUSTER 3000



SPAM Texts

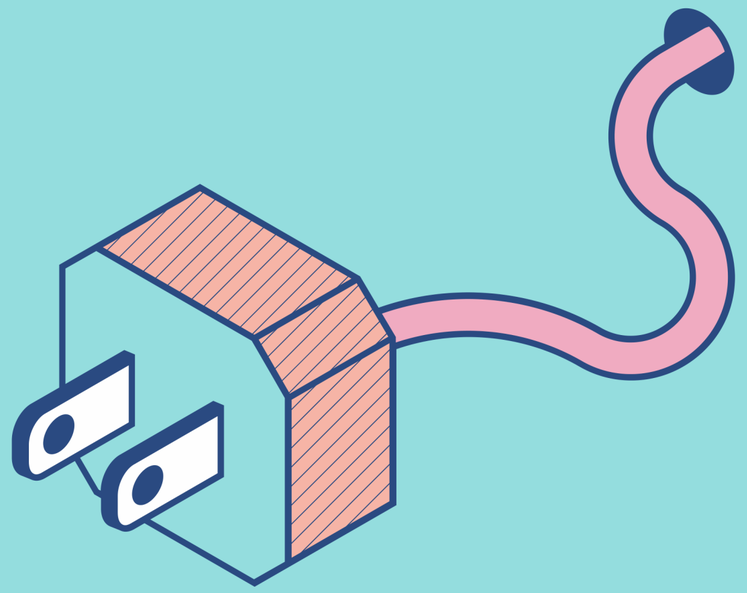
- Unsolicited, unwanted, or irrelevant messages that are sent to a large number of recipients via text message.
- Often contain advertisements, scams, or other types of unwanted content
- Phishing attempts could result in loss of money or spreading of fake news.

HAM Texts

- Ham texts refer to non-spam messages that are sent without malicious intent and are generally used to communicate between parties

Problem Definition

How can we effectively identify Spam messages with the attributes of text messages?

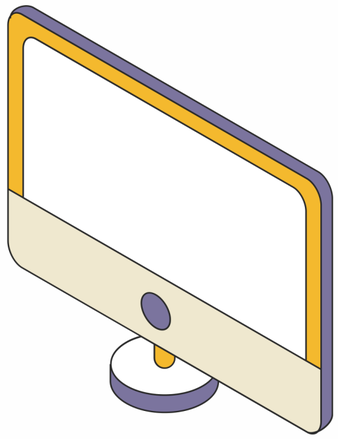


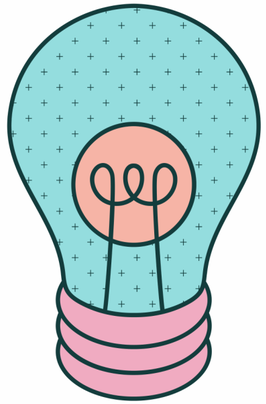
Dataset

spam.csv from Kaggle

- v1: a label that marks the nature of the message (spam/ham)
- v2: the text of the message
- Unnamed: 2 - 4: parts of original message that have been split

Contained 4825 Ham and 747 Spam





OVERVIEW

1 STEP

DATA CLEANING

2 STEP

DATA PREPARATION

3 STEP

EXPLORATARY DATA ANALYSIS

4 STEP

MODEL TRAINING

5 STEP

INSIGHTS & CONCLUSION

Step 1: Data Cleaning

1. Concatenating of Unnamed: 2 - 4 and 'v2' to form complete text message
2. Renaming columns 'v1' to 'Category' and 'v2' to 'Message'
3. Removal of duplicates

In [4]:
```
'full_message'] = df['v2'].fillna('') + ' ' + df['Unnamed: 2'].fillna('')+ ' ' + df['Unnamed: 3'].fillna('')+ ' ' +
```

In [5]:
```
df.rename(columns={'v1':'Category', 'full_message':'Message'}, inplace = True)
df = df.drop(['v2','Unnamed: 2', 'Unnamed: 3','Unnamed: 4'], axis=1)
```

In [6]:
```
df.head()
```

Out[6]:

	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...

Phone Number Detection

In [13]:
```
pattern = r'\d{6,}' ## Detect messages containing a strin

## Create a new column to store the results of the patter
df['Contain_HP'] = df['Message'].apply(lambda x: bool(re.
df.head(10)
```

Out[13]:

	Category	Message	Contain_HP
0	0	Go until jurong point, crazy.. Available only ...	False
1	0	Ok lar... Joking wif u oni...	False
3	0	U dun say so early hor... U c already then say...	False
4	0	Nah I don't think he goes to usf, he lives aro...	False
5	1	FreeMsg Hey there darling it's been 3 week's n...	False
6	0	Even my brother is not like to speak with me. ...	False
10	0	I'm gonna be home soon and i don't want to tal...	False
13	0	I've been searching for the right words to tha...	False
14	0	I HAVE A DATE ON SUNDAY WITH WILL!!	False
15	1	XXXMobileMovieClub: To use your credit, click ...	False

In [14]:
```
df['Contain_HP'].value_counts()
```

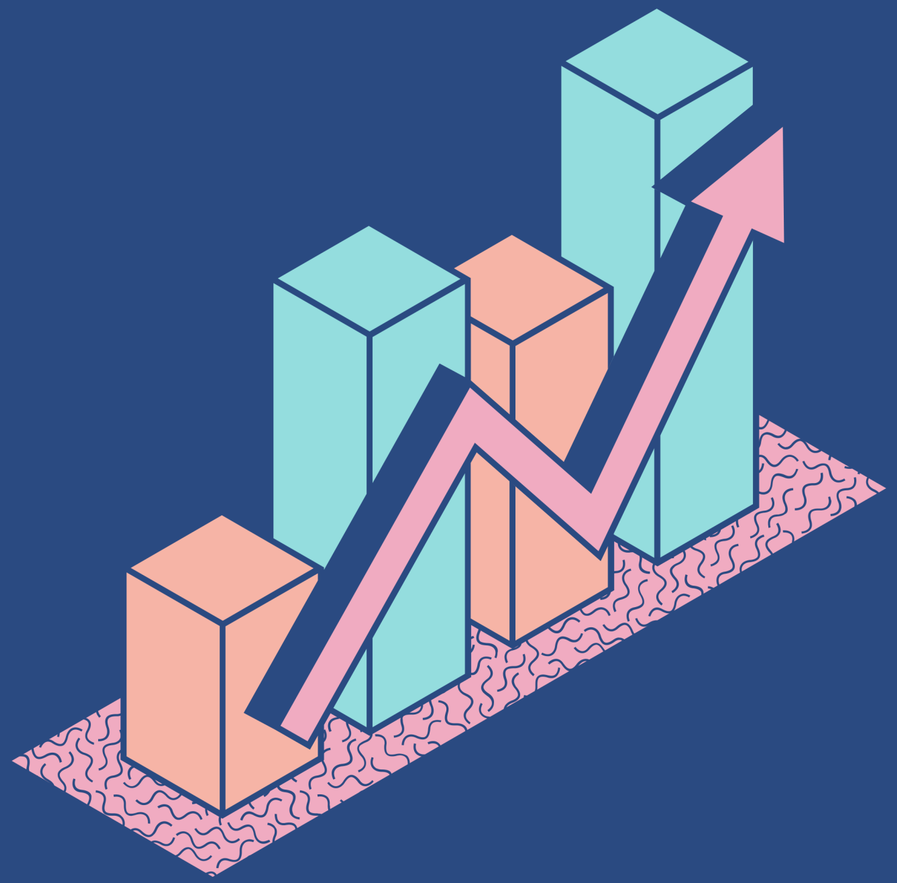
Out[14]:
```
False    4801
True      368
Name: Contain_HP, dtype: int64
```

Step 2: Data Preparation

- Generating stop words
  - Common words that can be ignored as they are unimportant.
  - Removal allows algorithm to focus on the important words instead.

- Creating boolean column Contain_HP
  - Indicates if text message contains a string of 6 or more numbers

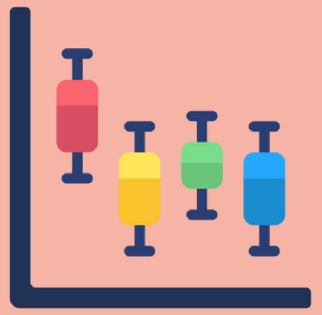
Step 3: EXPLORATORY ANALYSIS



Analysis Factors

Phone Number Analysis

Analysing how many Spam and Ham message contain possible phone numbers

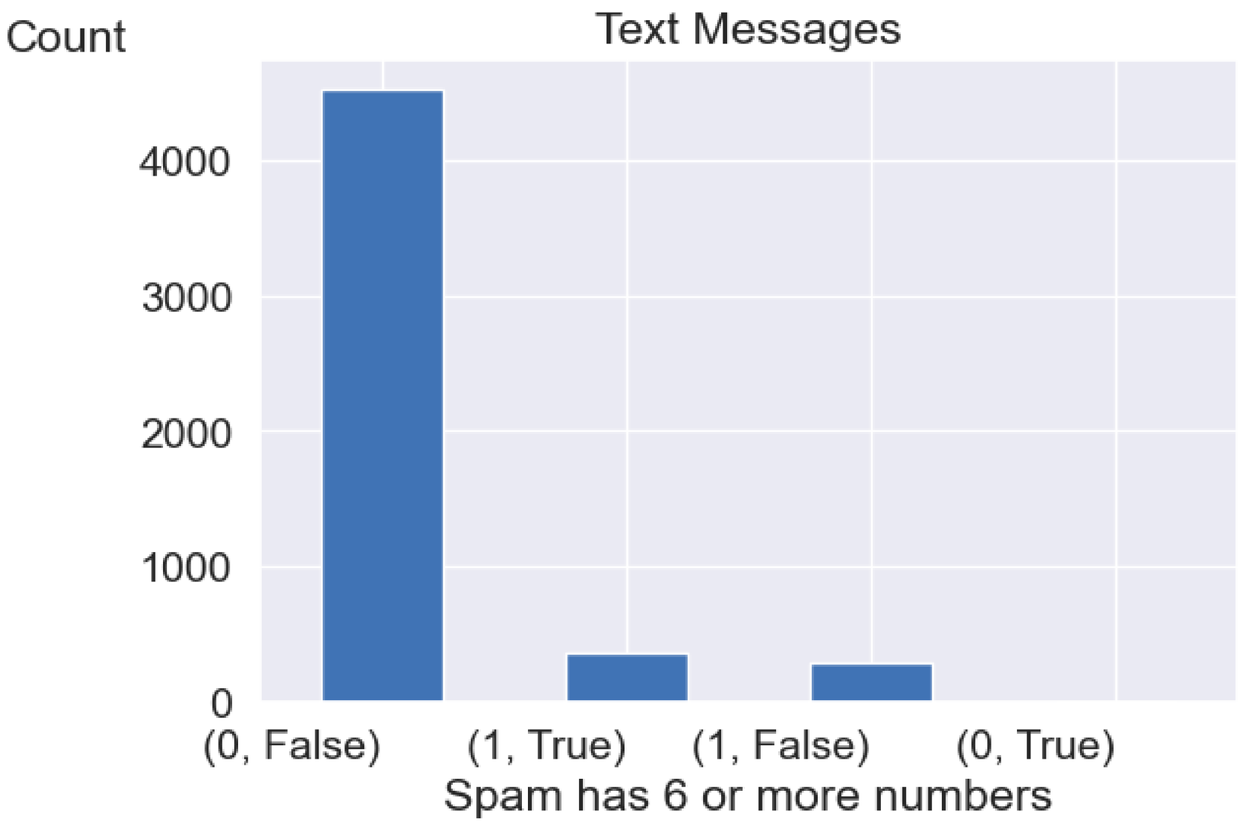


Word Count Analysis

Comparing word counts of Spam and Ham messages by understanding their distributions

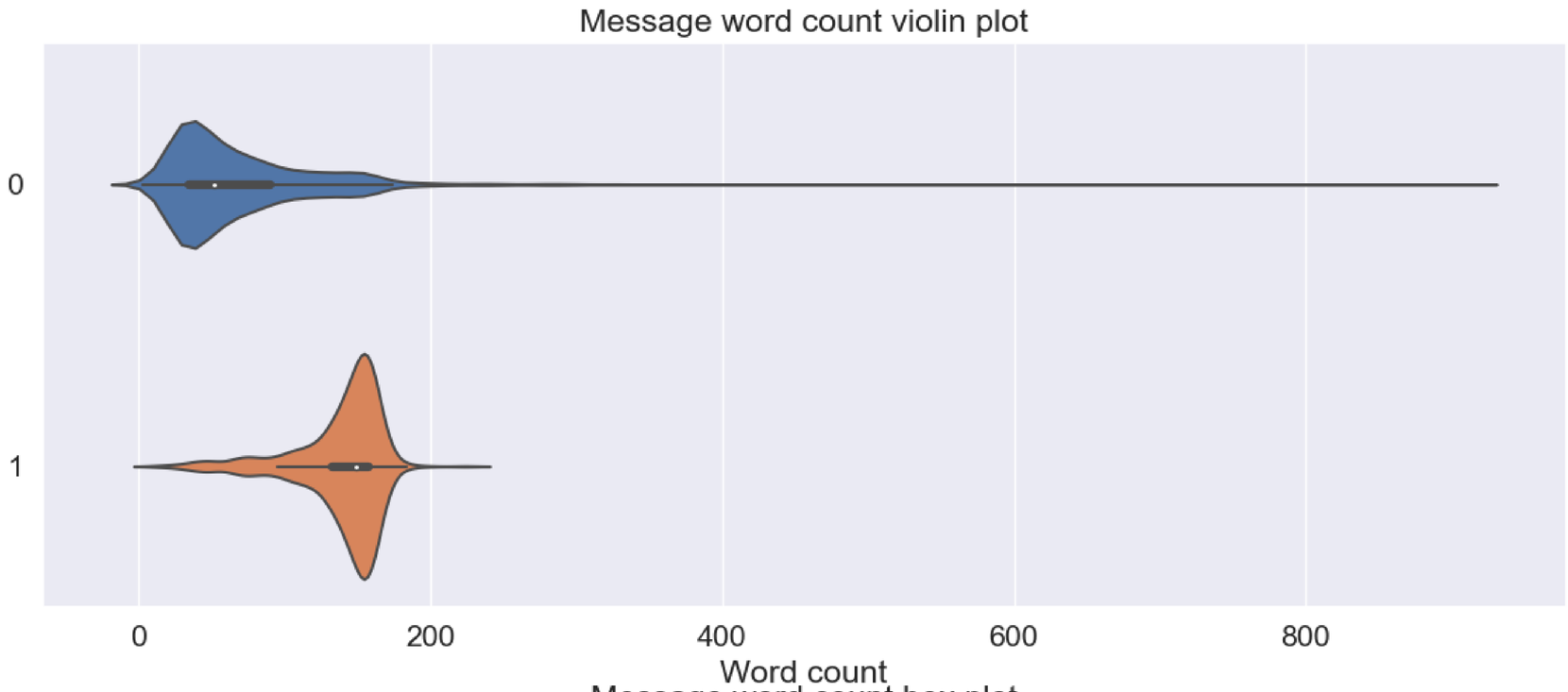
Most Frequent Words

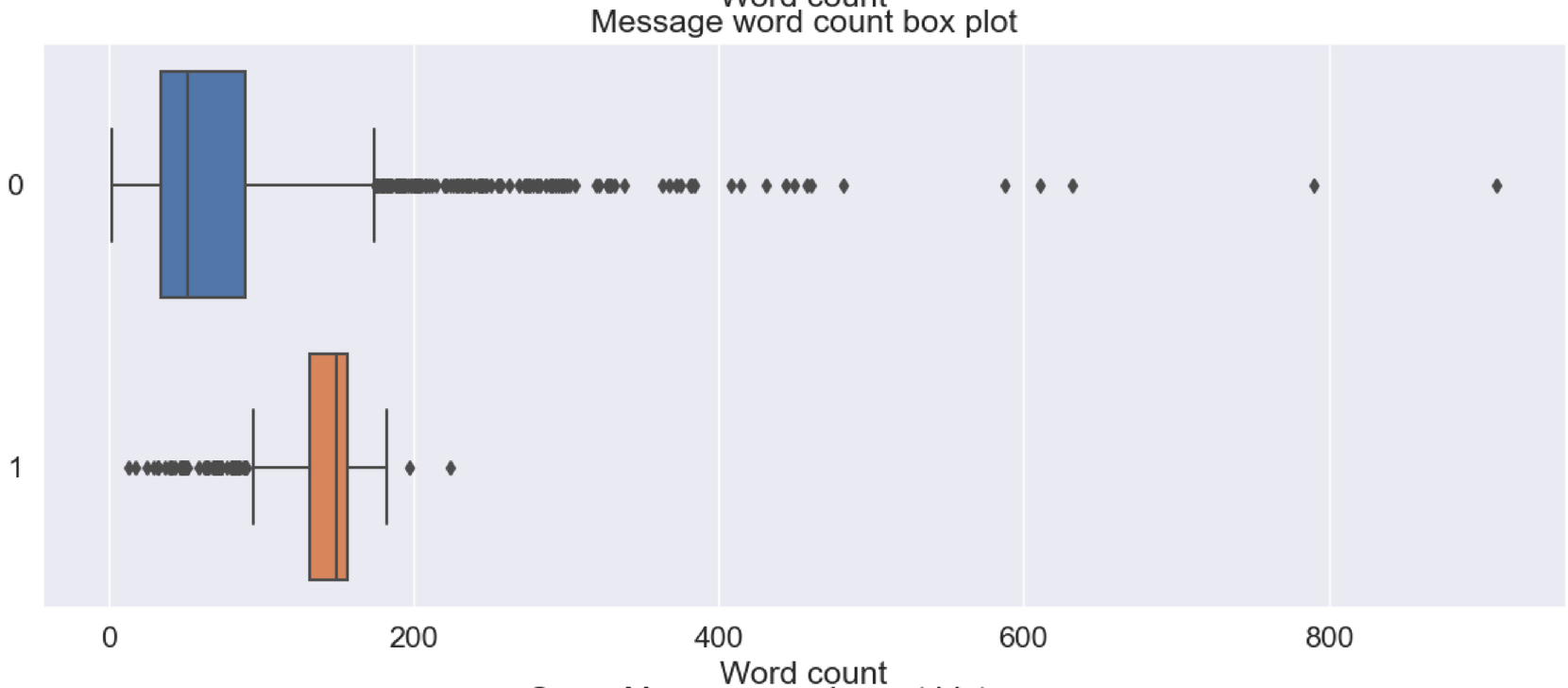
Analysing words that appear most in Spam and Ham messages through a bar graph and word clouds



Phone Number Analysis

- Most text messages that contain Phone Numbers are Spam messages

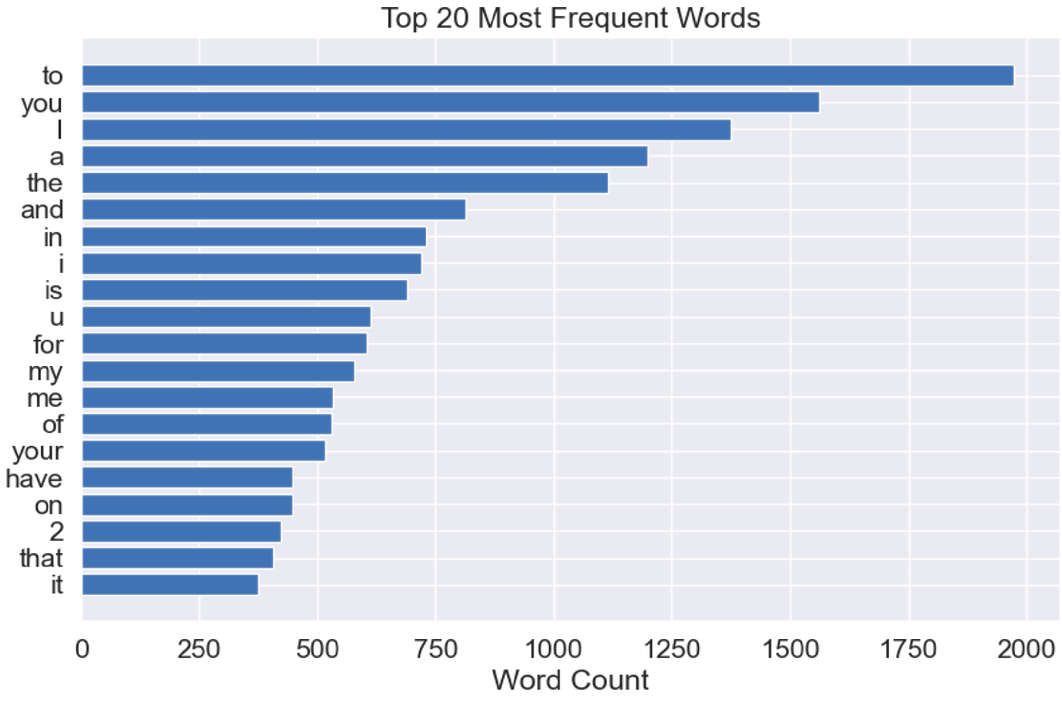




Word Count Analysis

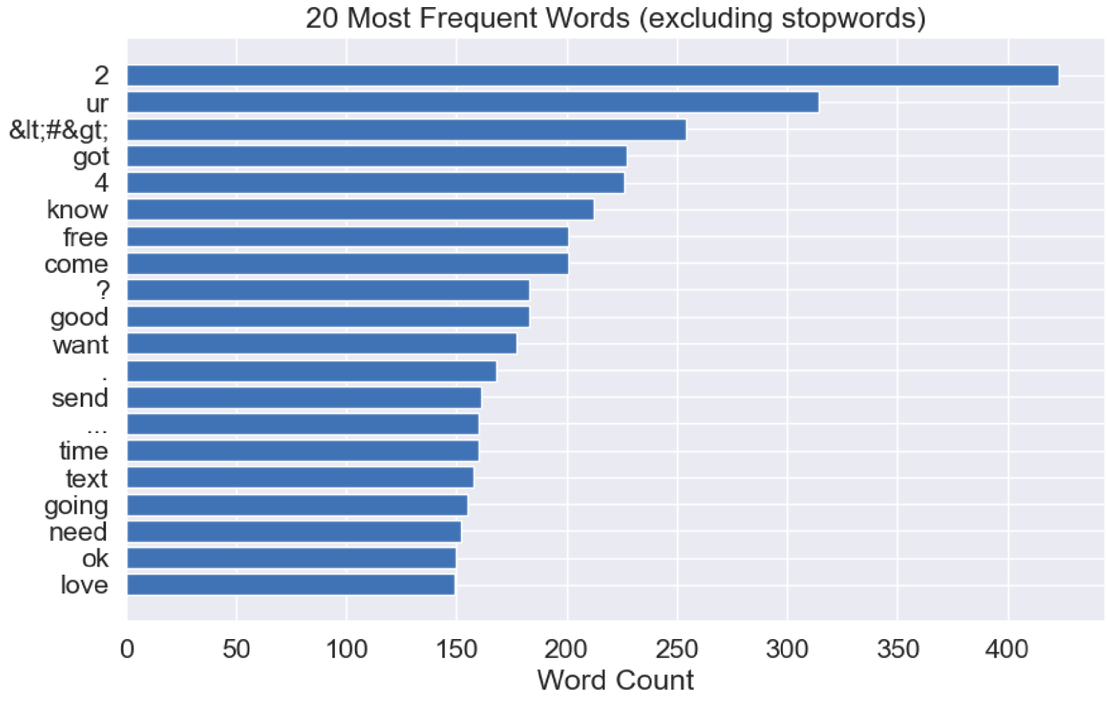
- Median number of words in SPAM messages > HAM messages
- Range of number of words in HAM messages > SPAM messages

Most frequent words in all messages



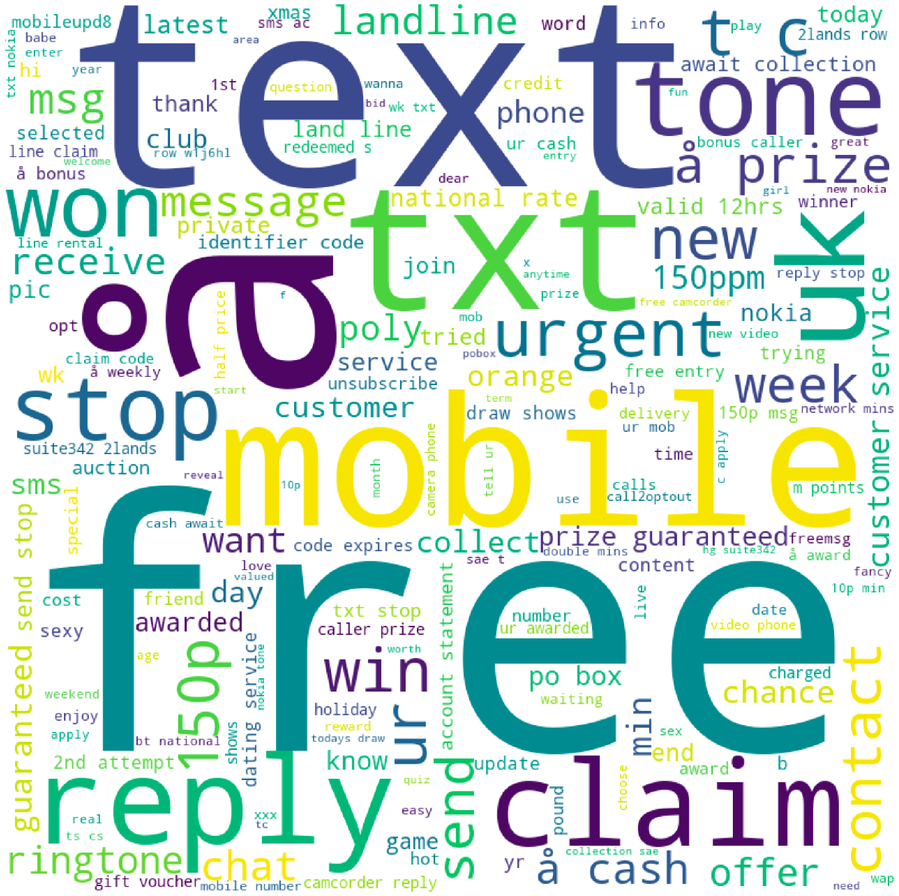
All Words

VS



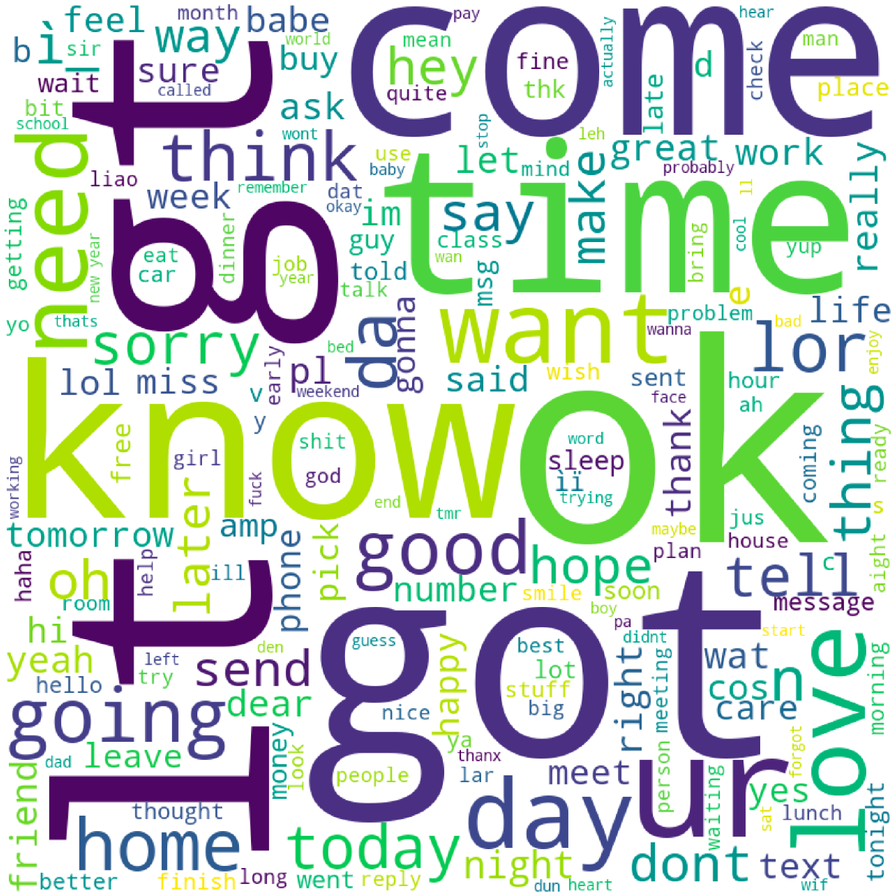
Stop Words Removed

Most Frequent Words



SPAM Messages (stop words removed)

vs


come
time
to
know
ok
got
lt
gt
love
ur
want
need
lor
thing
really
great
work
make
say
think
week
sorry
lol miss
good
hope
tell
number
soon
send
going
home
today
dont
text
night
dear
leave
friend
later
amp
phone
pick
tomorrow
oh
hi
yeah
hello
room
haha
meet
day
right
care
cos
wat
yes
lunch
tonight
morning
aight
ready
enjoy
s
coming
thank
sleep
ii
trying
sent
said
wish
hour
ah
life
lor
bad
wanna
problem
bring
cool
yup
ll
place
man
check
actually
hear
pay
d
late
fine
thk
mean
world
buy
hey
quite
babe
way
feel
month
bi
sir
wait
called
sure
bit
school
stop
wont
use
baby
okay
dat
remember
liao
week
eat
car
dinner
job
year
talk
told
guy
im
say
class
wan
msg
make
let
mind
great
work
probably
bring
cool
yup
really
ll
new year
getting
yo
thats
bed
sorry
lol miss
pl
weekend
early
v
y
da
gonna
go
said
wish
sent
face

Step 4: MODEL TRAINING

Models Used

Naive Bayes

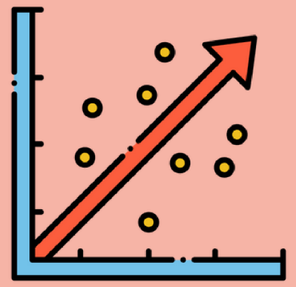
A classification algorithm based on weights assigned to individual words

Support Vector Machine

A prediction algorithm based on the features of the messages

Random Forest Classifier

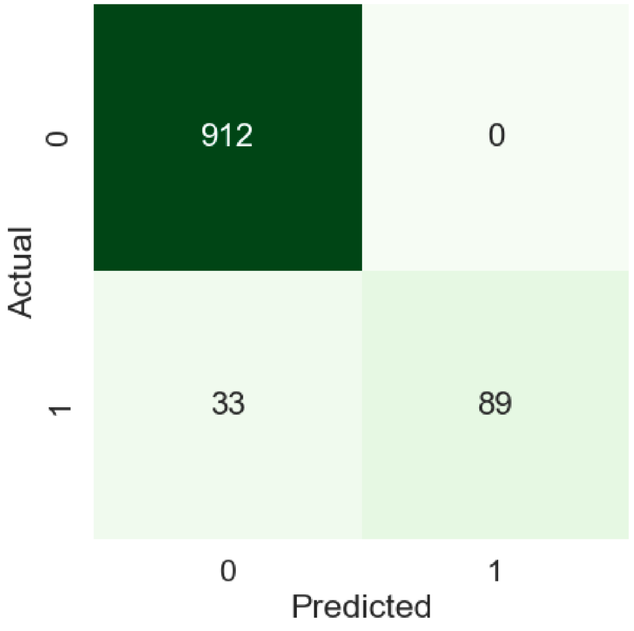
A classification algorithm based on multiple decision trees



Logistic Regression

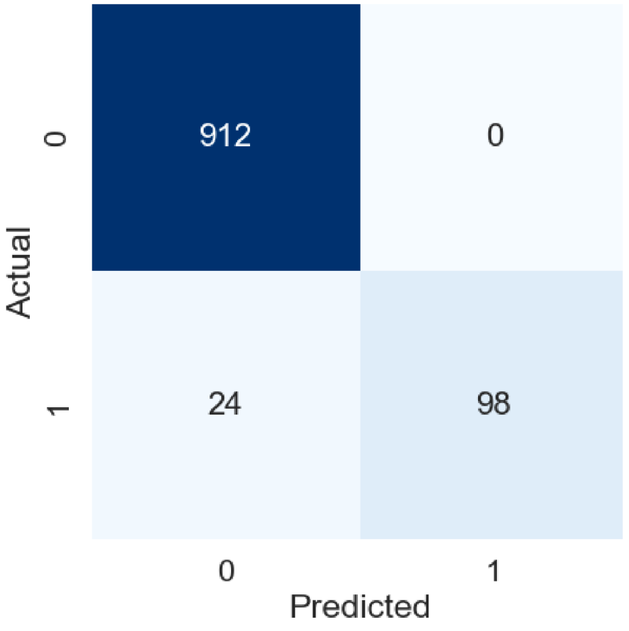
A classification algorithm that predicts based on 'Message' and 'Contain_HP'

Confusion Matrix



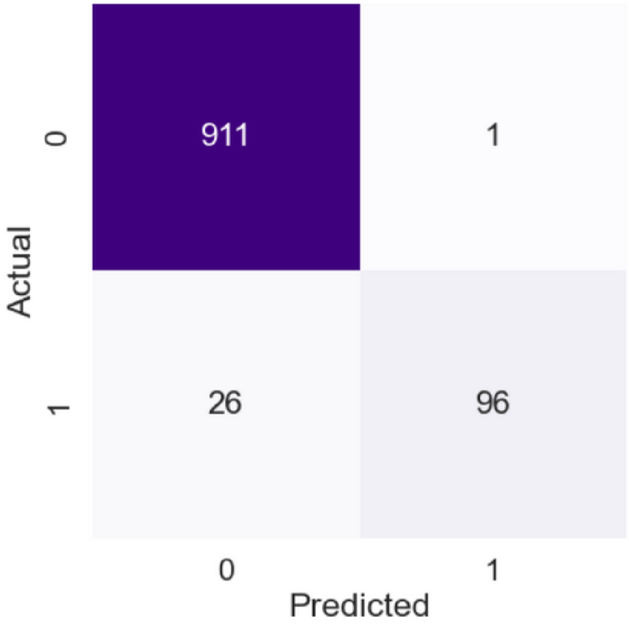
Naive Bayes

Accuracy: 0.968



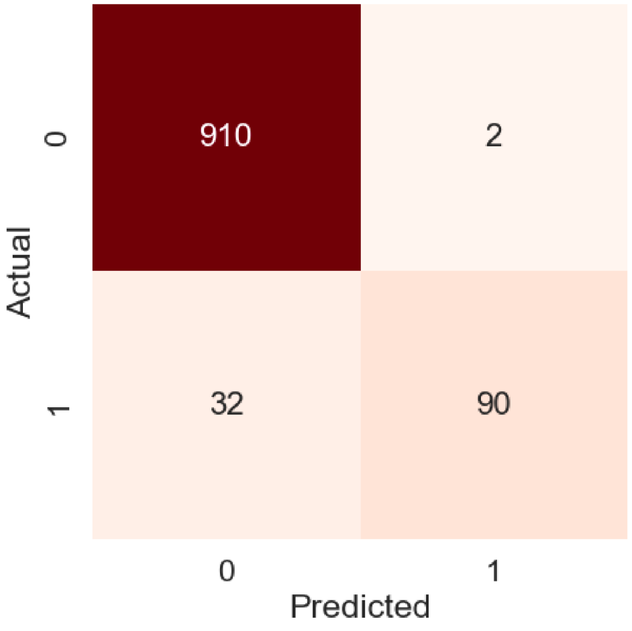
Support Vector Machine

Accuracy: 0.977



Random Forest Classifier

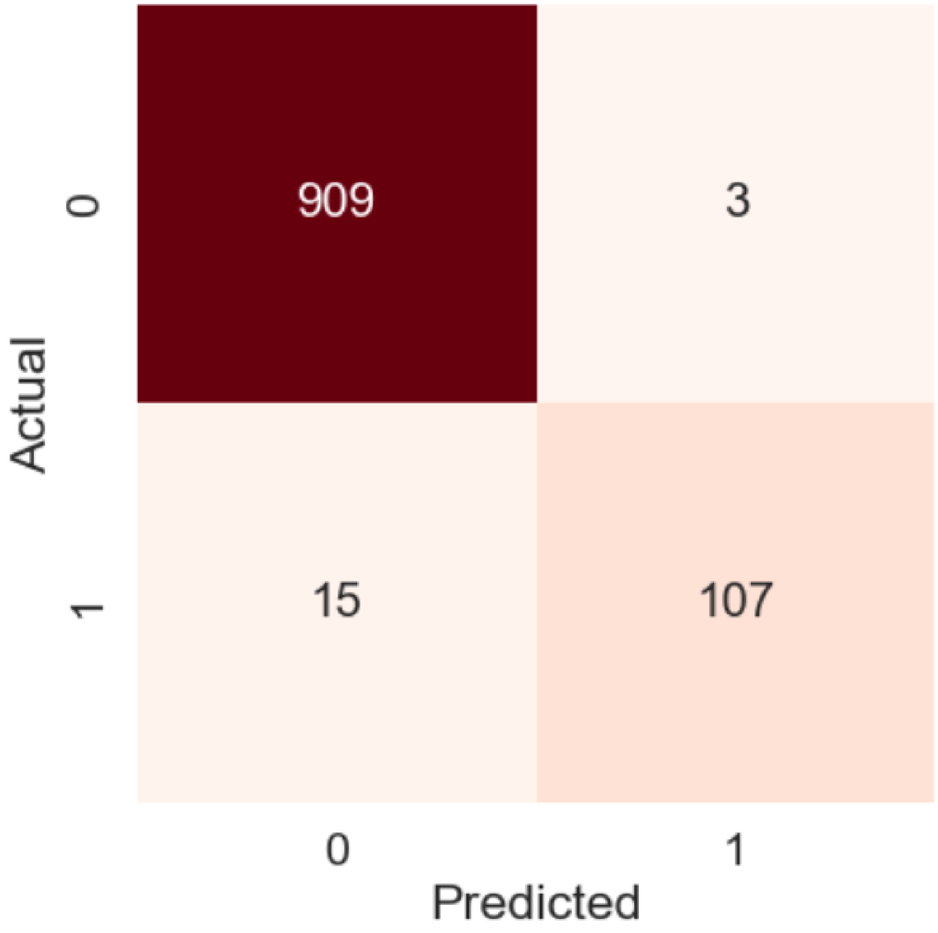
Accuracy: 0.978



Logistic Regression

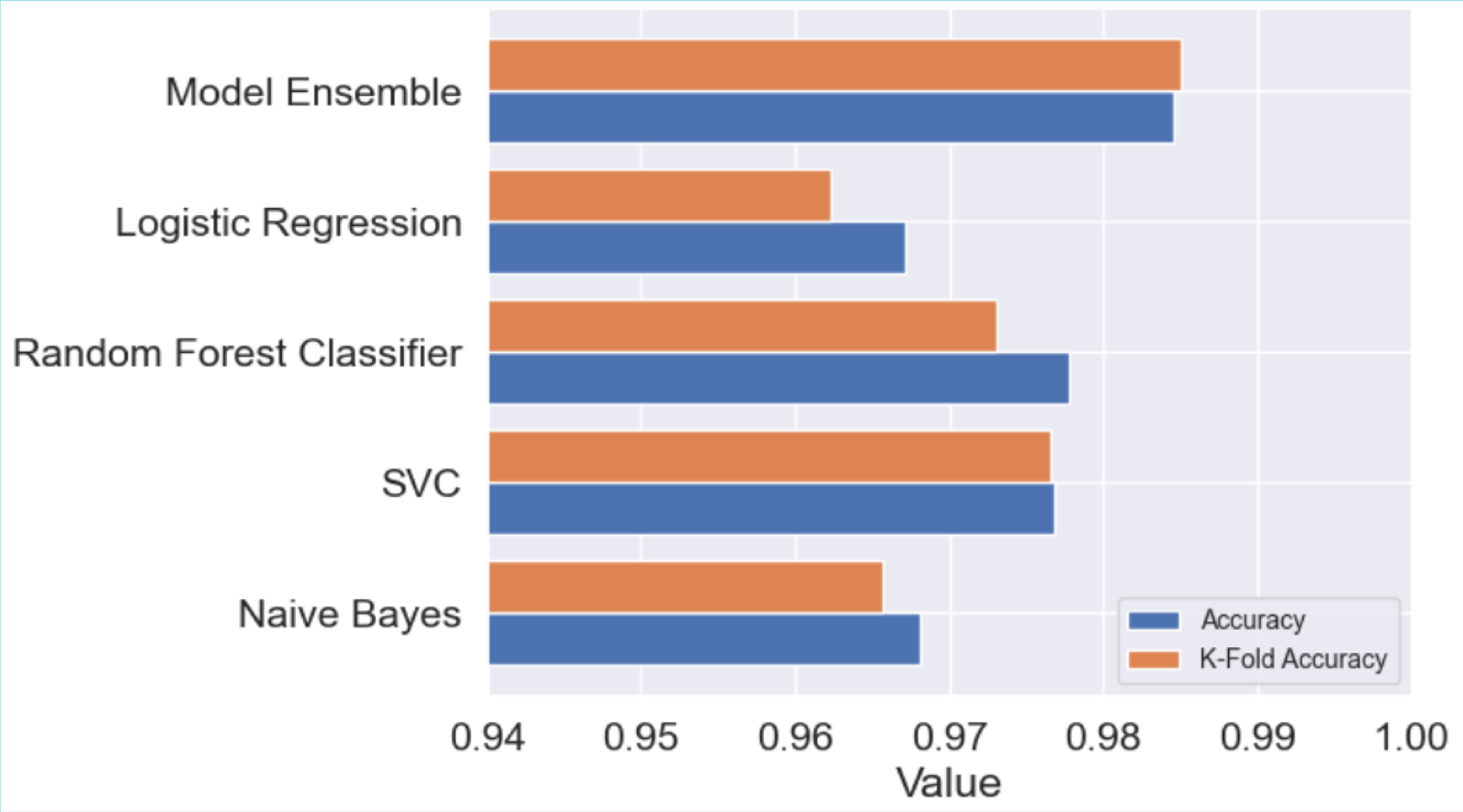
Accuracy: 0.967

Model Ensemble



- Combines results of 4 models
- Implements voting system
- Total score > 0 would be taken as positive

Accuracy: **0.985**



K-Fold

- Large disparity in number of SPAM and HAM in original data set
- More accurate measurement

Analysis of False Negatives

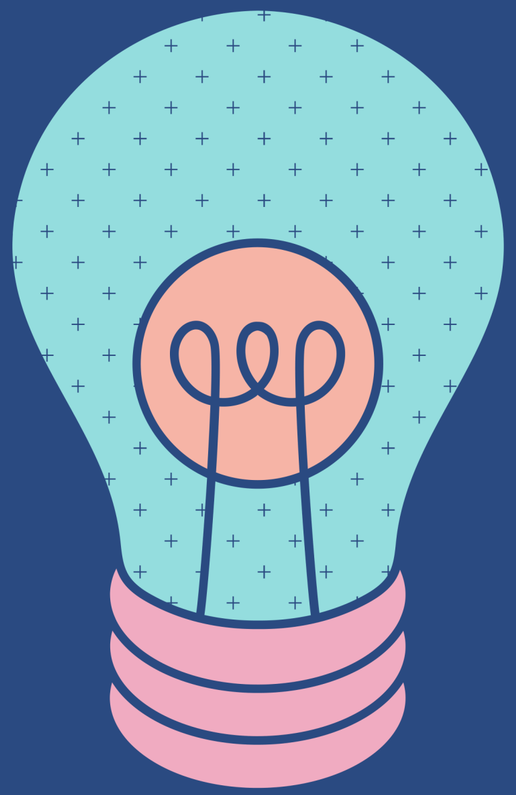
Total Number of False Negatives = **111**

Unique False Negatives = 50

Analysis of False Negatives

Number of Duplicated Messages = **61**

Step 5:
INSIGHTS & CONCLUSION

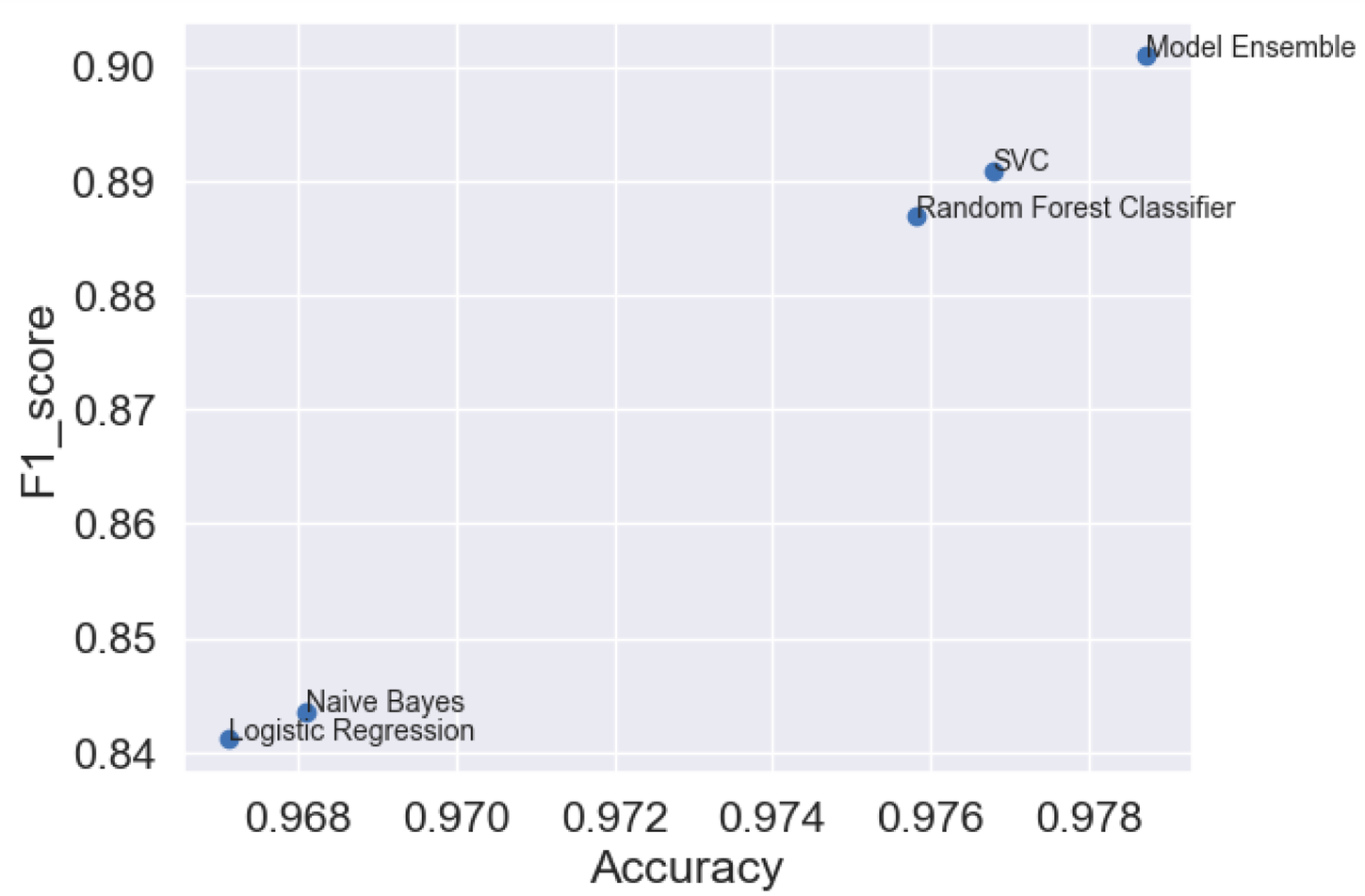


Consistent Errors

- Presence of misspellings
- Contain few spam-related words / attention grabbing words

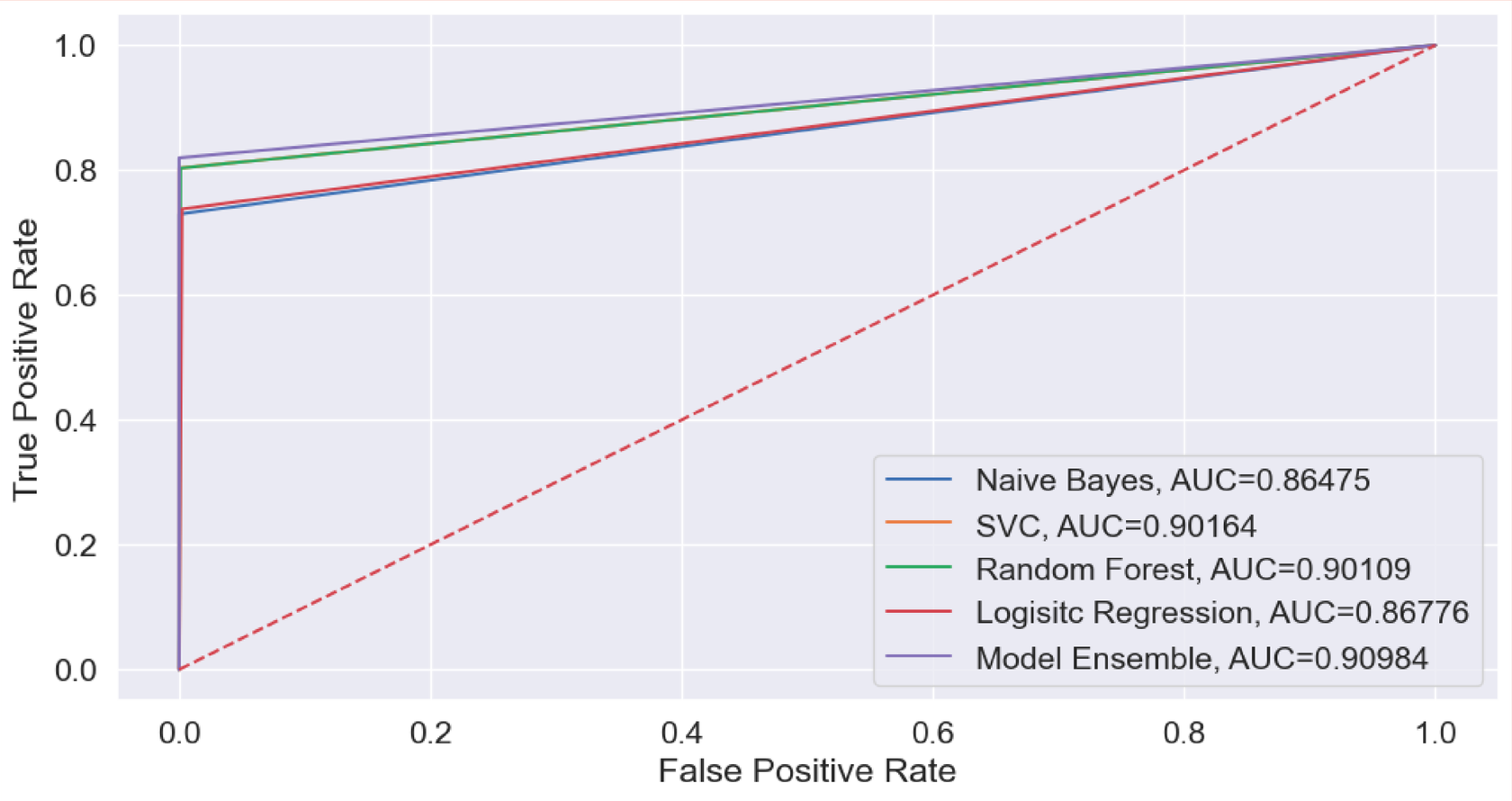
1 : message brought GMW Ltd. connected
2 : (Bank Granite issues Strong-Buy) EXPLOSIVE PICK MEMBERS *****UP 300% ********** Nasdaq Symbol CDGT $5.00 per..
3 : ringtoneking 84484
4 : Money won wining number 946 wot
5 : New Tones week include: 1)McFly-All Ab.., 2) Sara Jorge-Shock.. 3) Smith-Switch.. order follow instructions message
6 : Goal! Arsenal 4 (Henry, 7 v Liverpool 2 Henry scores simple shot 6 yards pass Bergkamp Arsenal 2 goal margin 78 mins.
7 : amazing xxx PICSFREE1 video sent enjoy! vid 2day text keyword PICSFREE1 video.
8 : Ur balance å£600. question: Complete landmark, Big, A. Bob, B. Barry C. Ben ?. Text A, B C 83738. Good luck!
9 : SMS. ac Sptv: New Jersey Devils Detroit Red Wings play Ice Hockey. Correct Incorrect? End? Reply END SPTV
10 : sale - arsenal dartboard. Good condition doubles trebles!
11 : 2/2 146tf150p
12 : Hi sue. 20 years old work lapdancer. love sex. Text live - bedroom now. text SUE 89555. TextOperator G2 1DA 150ppmsg 18+
13 : Fantasy Football TV. Sky Gamestar Sky Active play å£250k Dream Team. Scoring starts Saturday, register now!SKY OPT 88088
14 : Sorry! unsubscribe yet. MOB offer package min term 54 weeks> pls resubmit request expiry. Reply THEMOB HELP 4 info

Insights



Model Ensemble has highest Accuracy and F1 Score

ROC Curve



Model Ensemble is best classifier

Final product: SPAM BUSTER 3000!

```
 ____                    ___          __
/ __/___ ___ ____ __    / _ )__ _____/ /___ ____
\_ \/ _ \/ _ `/ _ `/ _ \  / _ / / / / __/ _/ _ \/ __/
__/ / /_/ / /_/ / / / / / / /_/ / /_/ (_ ) /_ _/ /
/___/ .__/\_,_/_/ /_/ /_/  /____/\_,_/___/\__/\__/_/
   /_/

                 ____ ___  ___  ___
                |__ // _ \/ _ \/ _ \
                 /_ </ / / / / / / /
                __/ / /_/ / /_/ / /_/ /
               /____/\___/\___/\___/

__  __          _          ____      ___      __
\ \/ /___  __  _( )_______  / ___/___ _/ /__  / /
 \  / __ \/ / / /// ___/ _ \ \__ \/ _ `/ /_ _ \/ /
 / / /_/ / /_/ / / /  /  __/ ___/ / /_/ / __/  __/_/
/_/\____/\__,_/ /_/   \___/ /____/\__,_/_/  \___(_)
```

- Takes in user input and determines whether it is Ham or Spam using Model Ensemble
- Integrates user feedback loops to train model that learns from real-world user interaction
- Increased accuracy with every prediction!

Future Improvements

- Training on larger and more balanced dataset
- Incorporation of deep neural networks machine learning techniques or natural language processing techniques

Thank you

SPAM BUSTER 3000 (enter 'quit' to exit):

Input Here