SPAM OR HAM CLASSIFICATION OF SMS TEXTS

Utsav Bhadresh Shah (ushah21@uic.edu)

Tuhin Kundu (tkundu2@uic.edu)

Sai Aakarsh Reddy Koppula (skoppu3@uic.edu)

Department of Computer Science

Introduction to Data Science - CS 418

Problem Statement

- *Congrats!! You just won a $1 million dollar. Fill in the below form with your details to get the money in your account.*
- How often have we come across such messages saying that we have won some prize or won a trip to Hawaii. This form of scam through sms which are generally spam texts is called smishing.
- The goal of our project is to use data science to build a model particularly text analysis and classification methods to classify the sms into spam or ham(non-spam).
- Ps: No Nigerian prince is sending you 1 million dollars.

Approach

- Step 1: Data Collection
- Step 2: Data Preparation
- Step 3: Data Exploration
- Step 4: Text Analysis
- Step 5: Classification

Data Collection

- SMS Spam Collection Data Set from UCI repository
- Data has two attributes which are, the sms texts and label which is either spam or ham
- The messages are a collection of personal text as well as a bunch of spam messages.
- There are a total of 5572 data points in this collection.

	class	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...

Data Preparation/Text Analysis

- There are a total of 5572 data points in the dataset.
- There are no null/missing values in the data. The data is imbalanced with 4825 instances of ham class and 747 instances of spam class.
- The original data and the processed data is shown below:

```
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
Even my brother is not like to speak with me. They treat me like aids patent.
As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. P
WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Cl
Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mob
```

```
freeman hey darl number week word back like fun still tb ok xxx std cha send moneysymb number rev
even brother like speak treat like aid patent
per request bell bell minnaminungint nurungu vietnam set callertun caller press number copi friend callertun
winner valu network custom select receiv moneysymb number prize reward claim call number claim code kl number valid numbe
mobil number month r entitl updat latest colour mobil camera free call mobil updat co free number
```

Data Preparation/Text Analysis

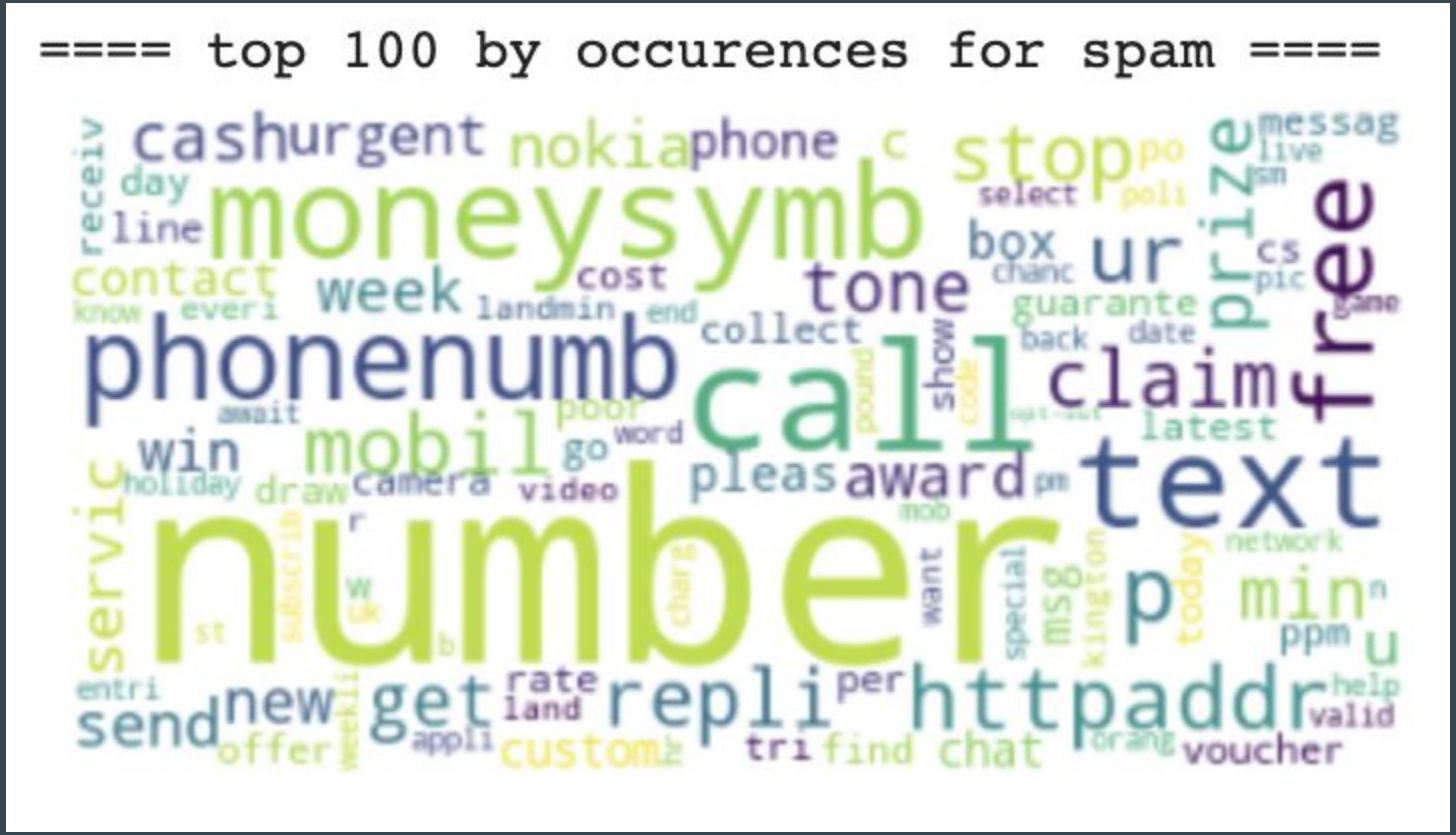
The steps involved are:

- Converted the messages to lowercase.
- Regex to substitute the email addresses, phone numbers, http links, symbols and numbers with a fixed string value.
- Regex to remove punctuations and replace them with whitespace.
- Implemented a spell corrector to take care of typos.
- Implemented a stemmer to stem the words.
- Removed the stopwords from the messages as they do not act as the differentiating factor.

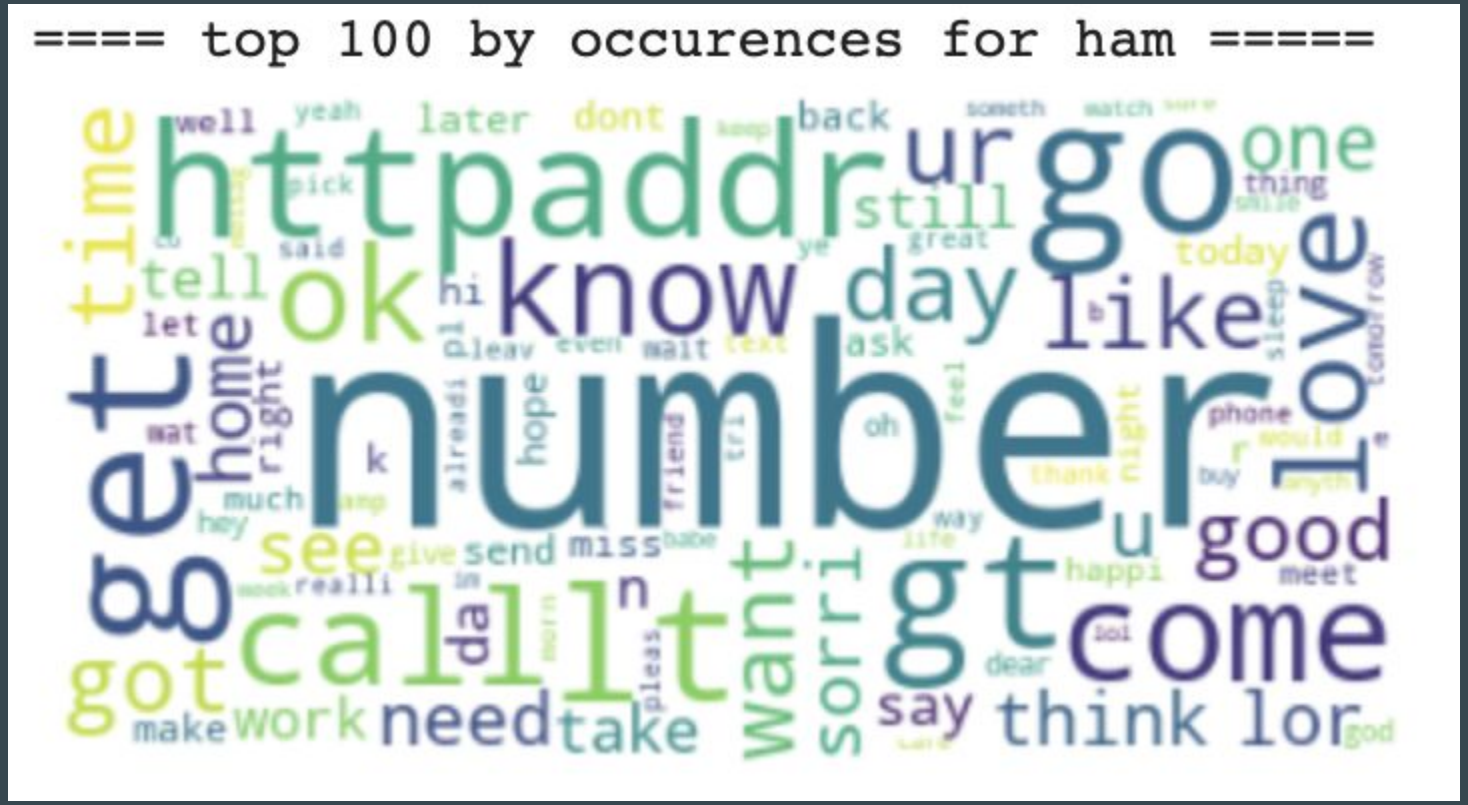
Data Preparation/Text Analysis

- Word count of the each token in the vocabulary was taken across all messages for exploration and visualization.
- The average number of tokens in each message was calculated for analysis.
- The length of the messages in the number of characters.
- TfIdf vectorizer was used to calculate the frequency of occurrence of words in each the sms messages based on the number of documents it occurs in.
- Count Vectorizer was also used for the data modeling exercise. It basically uses word frequency to create the vectors for each document(message).
- The TfIdf vectorized & Count Vectorized messages were then split into a train and test part and further used for the classification exercise.

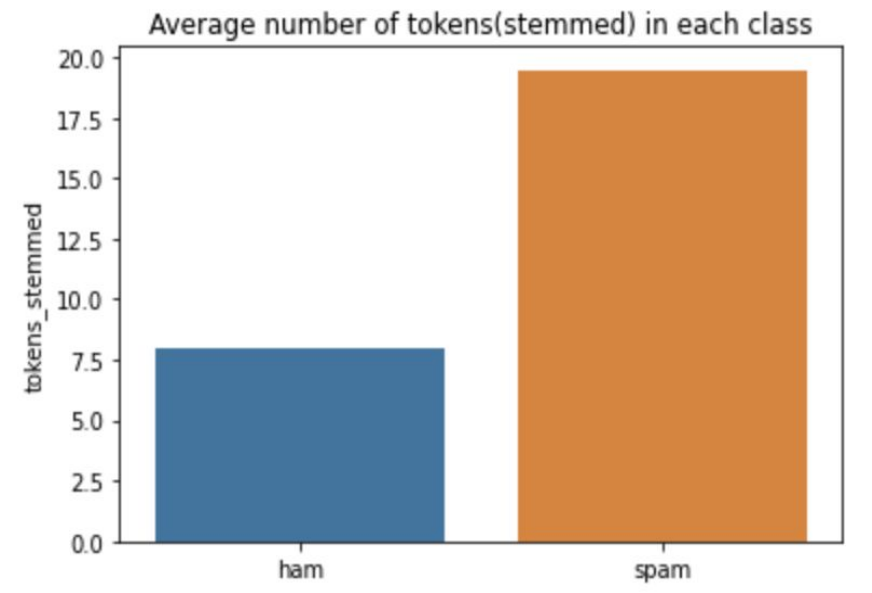
Data Exploration

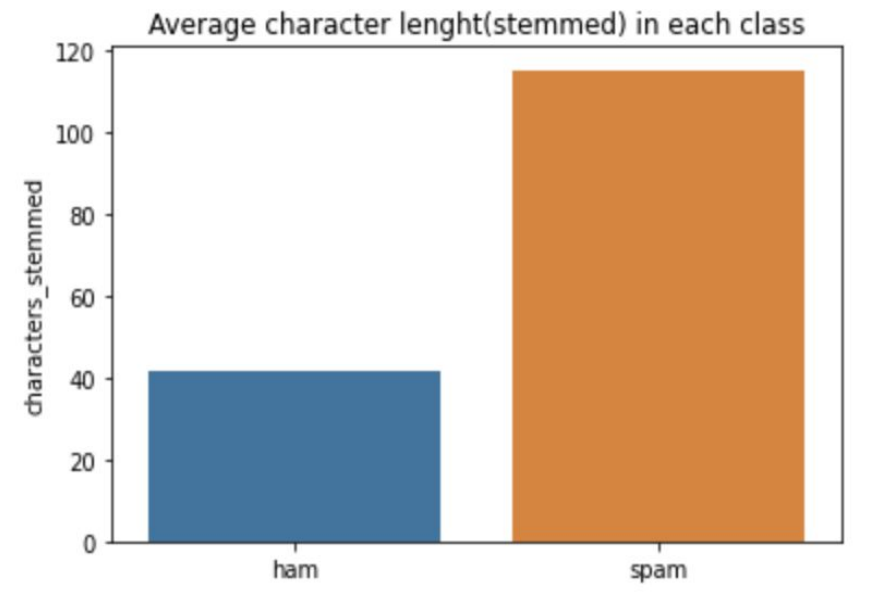


Data Exploration

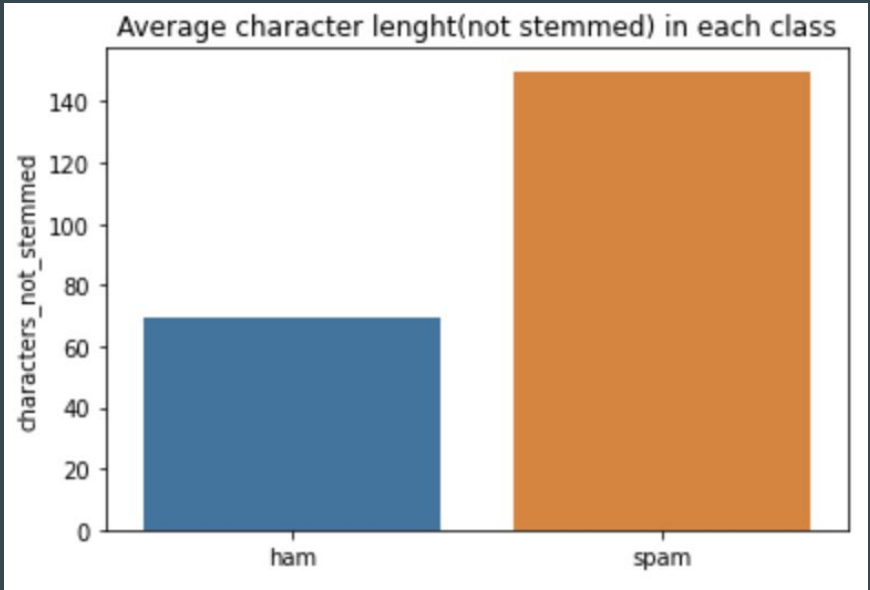


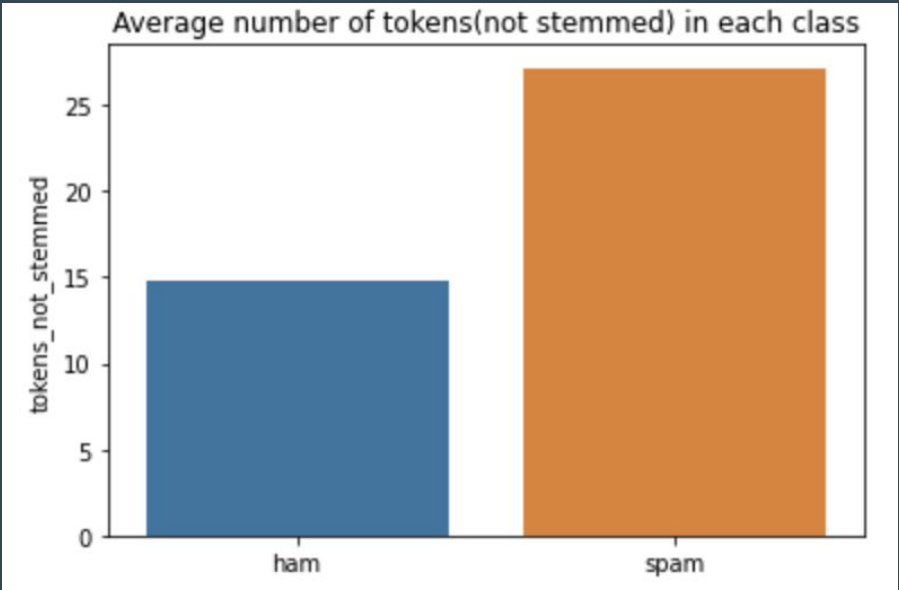
Data Exploration



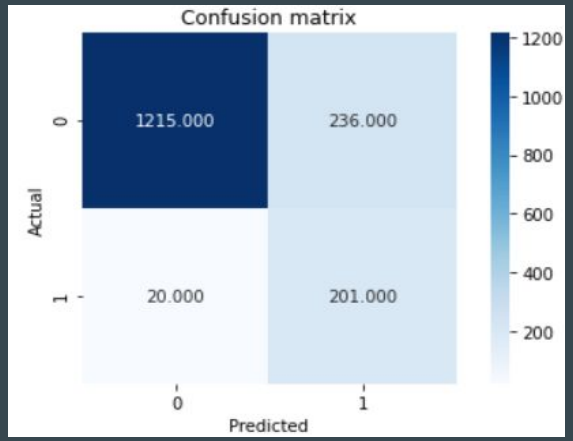


Data Exploration

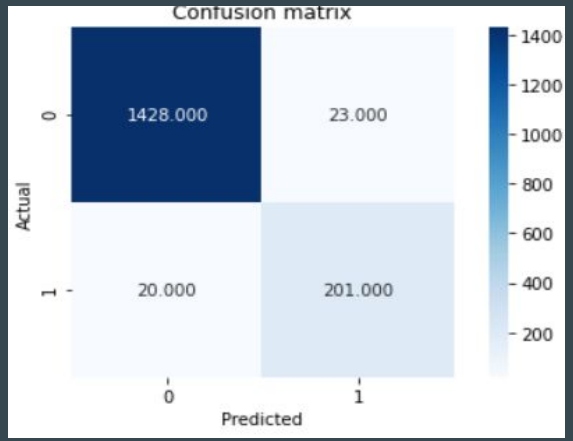




Classification models with frequency vectorizer

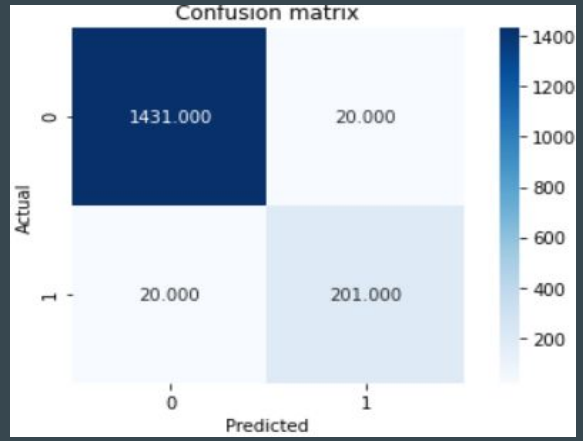


1) The model on the left is a gaussian naive bayes. With an accuracy of 84.6% and f1 score of 0.90 and 0.61

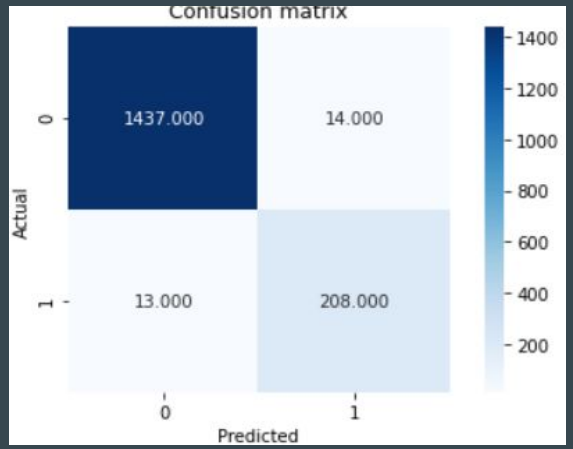


2) The model on the left is a decision tree classifier with entropy criterion. With an accuracy of 97.4% and f1 score of 0.98 and 0.90

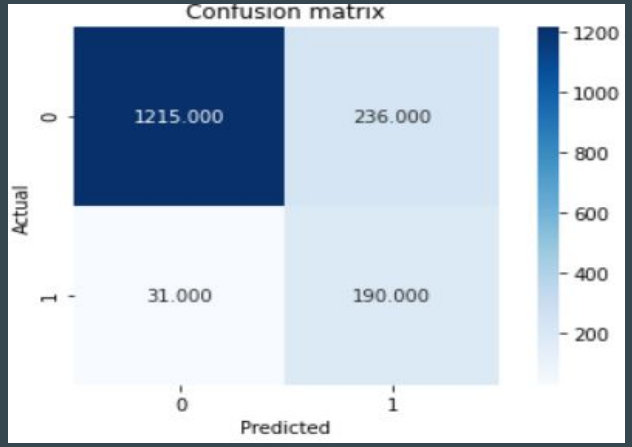
3) The model on the right is a decision tree classifier with gini criterion. With an accuracy of 97.6% and f1 score of 0.98 and 0.90



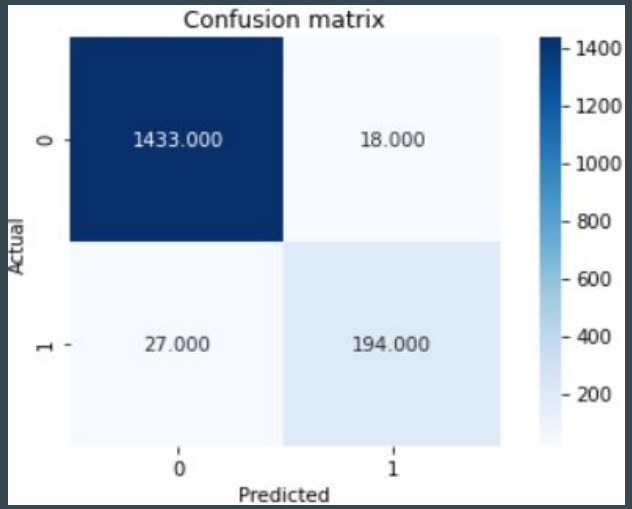
4) The model on the right is a SVM classifier. With an accuracy of 98.3% and f1 score of 0.99 and 0.93



Classification models with Tfidf vectorizer

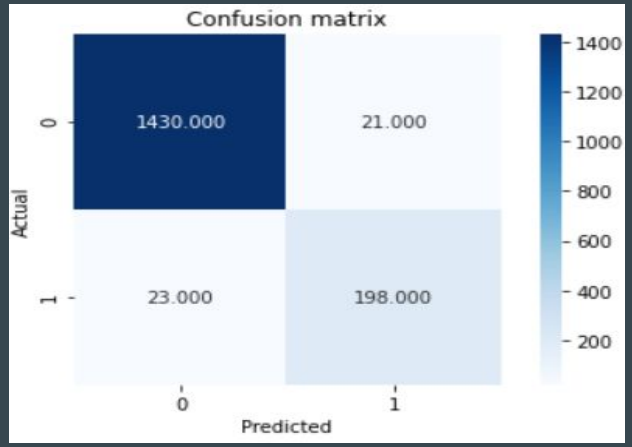


1) The model on the left is a gaussian naive bayes. With an accuracy of 84% and f1 score of 0.91 and 0.58

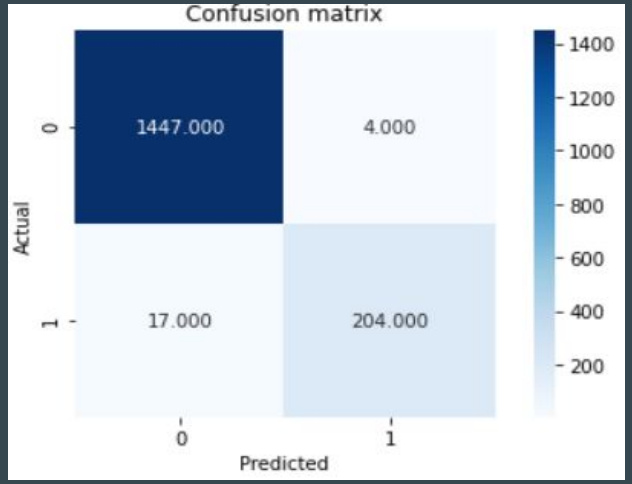


2) The model on the left is a decision tree classifier with entropy criterion. With an accuracy of 97% and f1 score of 0.98 and 0.89

3) The model on the right is a decision tree classifier with gini criterion. With an accuracy of 97.3% and f1 score of 0.98 and 0.90



4) The model on the right is a SVM classifier. With an accuracy of 98.7% and f1 score of 0.99 and 0.95



Classification results

* The below results are for models with the use of frequency vector.

Model	Accuracy	F1 score (ham)	F1 score (spam)
Naive Bayes	84.68	90.46	61.09
Decision Tree using entropy	97.42	98.51	90.33
Decision Tree using gini	97.60	98.62	90.95
SVM with linear	**98.38**	**99.06**	**93.90**

Classification results

* The below results are for models with the use of Tfidf vectorizer

Model	Accuracy	F1 score (ham)	F1 score (spam)
Naive Bayes	84.03	91.10	58.73
Decision Tree using entropy	97.30	98.45	89.60
Decision Tree using gini	97.36	98.48	90
SVM with linear	**98.74**	**99.27**	**95.10**


SUCH THANKS
WOW
SUCH
AMAZE
WOW
SO
APPRECIATE
VERY
GRATEFUL