Spam detection using Naïve Bayes Classification Algorithm

Probability Models (BANA 7031)
Final Project

Ashish Saxena

Index

Introduction to Bayes Theorem

- Named after Thomas Bayes
- Describes the probability of occurrence of an event, based on prior conditions that might be related to the event

Thomas Bayes
(Source: BBC.com)

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad , P(B) \neq 0$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad , P(A) \neq 0$$

- **P(A|B)**: Probability of occurrence of event A given event B has already occurred (*posterior probability*)
- **P(B|A)** : Probability of occurrence of event B given event A has already occurred
- **P(A)**: Probability of occurrence of event A
- **P(B)**: Probability of occurrence of event B

Introduction to Bayes Theorem

$$P(A|B) = \frac{P(B|A).\,P(A)}{P(B)}$$

What does it mean?

- If, for two events, conditional probability for any one event over other is defined, the conditional probability of the second over the first can be calculated given marginal probabilities of each event

How is it useful?

- It can helpful in categorizing new elements into specific buckets based on information from existing elements in each bucket

Introduction to Bayes Theorem

Predicting Play based on Weather Conditions of past two weeks

Weather	**Play**
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Weather	**Play - No**	**Play - Yes**		
Overcast		4	= 4/14	0.29
Rainy	3	2	= 5/14	0.36
Sunny	2	3	= 5/14	0.36
All	**5**	**9**		
	= 5/14	= 9/14		
	0.36	0.64		

$$\mathrm{P(Play - Yes|Sunny)} = \frac{\mathrm{P(Sunny|Play - Yes).P(Play - Yes)}}{\mathrm{P(Sunny)}}$$

$$\mathrm{P(Play - Yes|Sunny)} = \frac{(3/9).(9/14)}{5/14}$$

$$= \frac{3}{5}$$

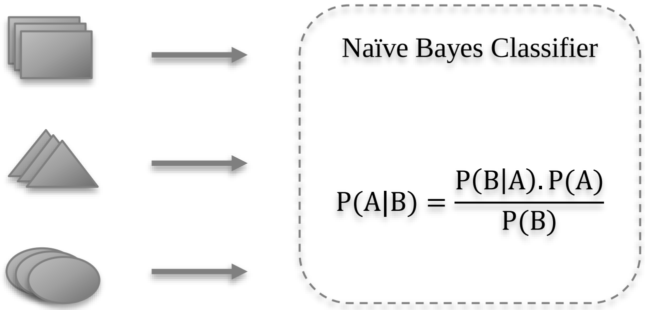
$$\mathrm{P(Play - No|Sunny)} = \frac{(2/5).(5/14)}{5/14}$$

$$= \frac{2}{5}$$

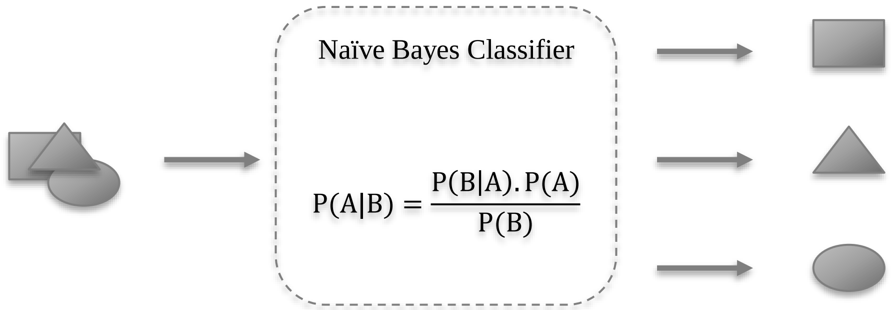
Naïve Bayes Classifier – Under the Hood

- Works on the underlying principle of Bayes Theorem
- Helps predict probability of an element belonging to a category basis its attributes

Training



Testing



$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Assumptions

- All predictor variables are independent and do not impact occurrence of each other
- All predictor variables hold equal importance in the prediction of response variable

Naïve Bayes Classifier – Under the Hood

PROS

- Simple, fast, and very effective
- Does well with noisy and missing data
- Requires relatively few examples for training, but also works well with very large numbers of examples
- Easy to obtain the estimated probability for a prediction

CONS

- Relies on an often-faulty assumption of equally important and independent features
- Not ideal for datasets with many numeric features
- Estimated probabilities are less reliable than the predicted classes

Problem Statement

- Telecom company J-Mobile has a base of over 50 million customers who receive spam messages on a regular basis from marketing teams of several companies
- These messages are sent from business accounts for which the company charges a higher tariff value to its advertisers for each sent message
- The company has recently noticed spam messages being sent from individual mobile numbers and wants to flag such messages sent to its customers on its own network to monetize better

Problem Definition

Current State

- J-Mobile is a telecom company with 50 million customers
- It charges certain specific amount for spam messages sent through its network
- The company wants to identify spam messages sent to customers to monetize better

Gap

J-Mobile does not have a spam detection framework in place

Key Question

How are spam messages different from non-spam messages?

Desired State

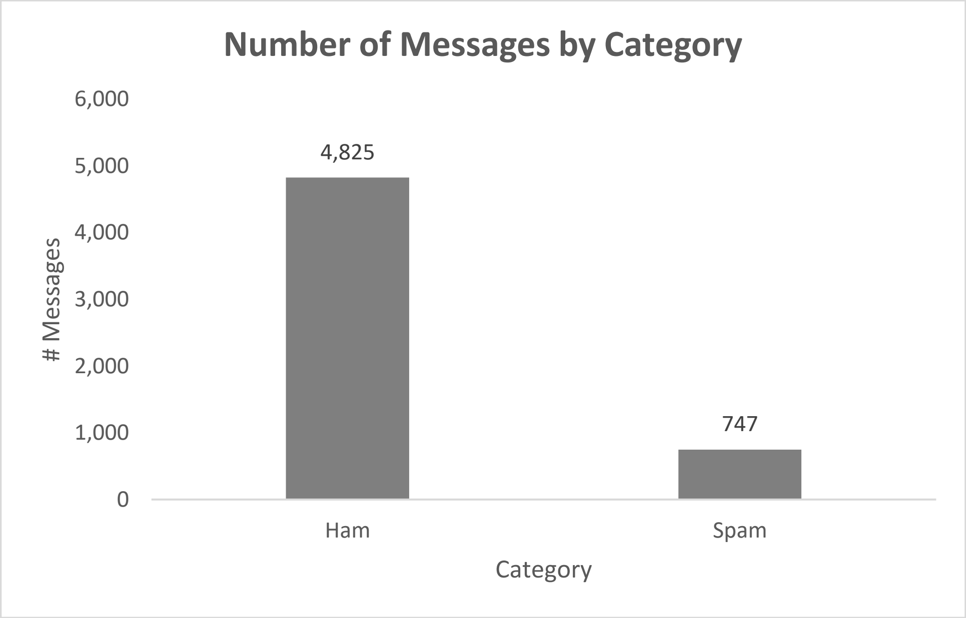
- **Outcome:** J-Mobile is able to achieve its monetization targets
- **Behavior:** J-Mobile is able to implement desired monetization strategy
- **Insight:** J-Mobile was able to identify spam messages sent to its customers

Executive Summary

- The created spam detection model had an accuracy of 98.15%
- The model exhibited a specificity of 89.62% thereby detecting correctly approximately 89 of every 100 spam messages
- The sensitivity of the model was observed to be 99.38% implying accurate detection of almost all non-spam messages
- The AUC value of the classification model was observed to be 0.9871 indicating high distinguishing capability between the classes

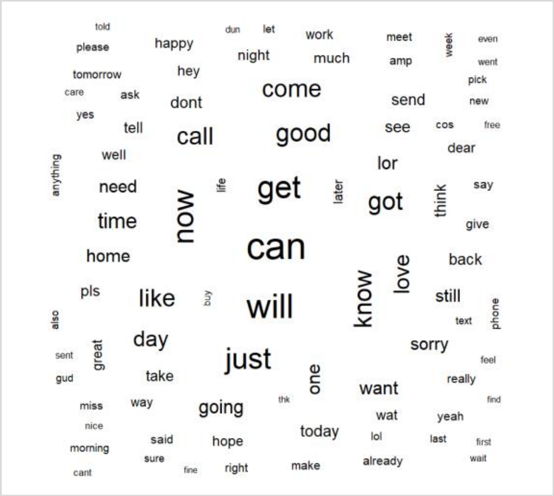
Exploratory Data Analysis & Results

Over 13% of the messages sent to customers were spam messages

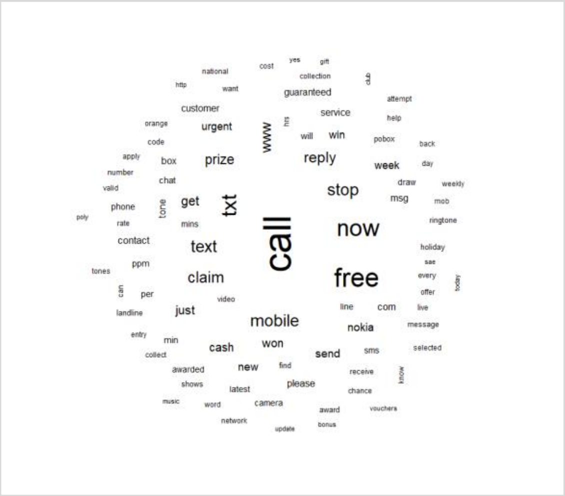


- Of the 5,572 messages in the dataset, 747 messages were observed as spam messages
- The non-spam (ham) messages were observed to constitute over 86% of the total messages

Frequent keywords across spam and non-spam messages differed significantly with few common keywords



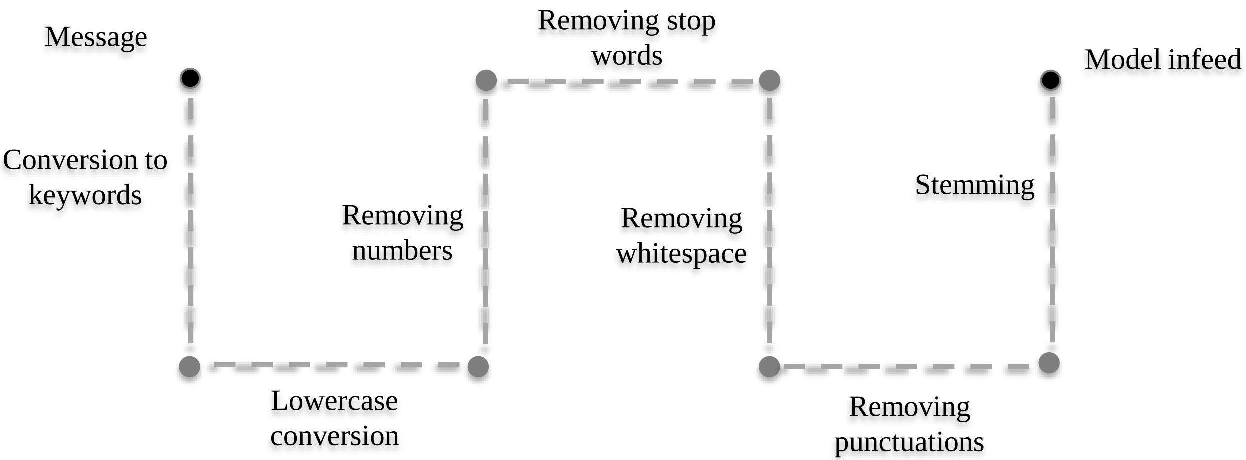
Ham (Non-Spam)



Spam

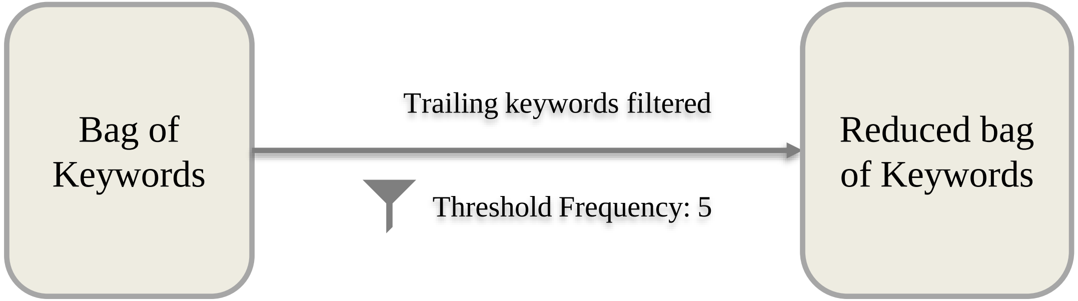
- While keywords across spam and non-spam messages differed significantly, common keywords included words like ‘call’, ‘now’ with varied frequencies across the two

Text messages received were processed for model infeed by removing numerical predictors, whitespace and implementing keyword stemming



- While most of the pre-processing involved basic text mining operations, the final step used stemming
  - Stemming is used to associate different forms of a verb to a single root verb
  - For example, ‘running’, ‘ran’, ‘runs’ are converted to the root verb ‘run’

Keywords occurring below a specific frequency threshold were removed to enable effective model training simultaneously reducing training time



- The keyword pool was filtered for a threshold frequency thereby reducing the number for training and testing datasets
  - This would lead to effective model training with reduced training time

A Document-Term matrix of the obtained keywords revealed the existence of keywords along with their prior categorization

Terms

Documents

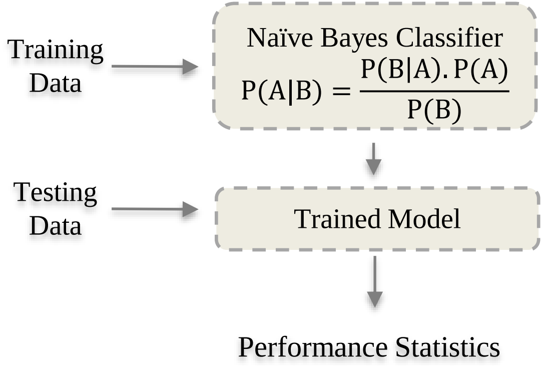
	KW_1	KW_2	KW_3	KW_4	KW_5	...	KW_n
Msg 1	Yes	No	Yes	Yes	No	...	Yes
Msg 2	No	No	Yes	No	Yes	...	No
Msg 3	Yes	No	Yes	Yes	No	...	Yes
Msg 4	No	Yes	No	Yes	Yes	...	Yes

Keywords

Messages

- The obtained terms (keywords) in each document (message) were converted to a document-term matrix for model training
- Since Naïve Bayes classifier works on categorical data, all frequencies of the occurrence were converted into their categorical equivalents (“Yes” or “No) based on presence

The Naïve Bayes classifier trained on the obtained data and further tested revealed encouraging performance statistics



Performance Statistics

	Reference (Ham)	Reference (Spam)
Predicted (Ham)	1,451	22
Predicted (Spam)	9	190

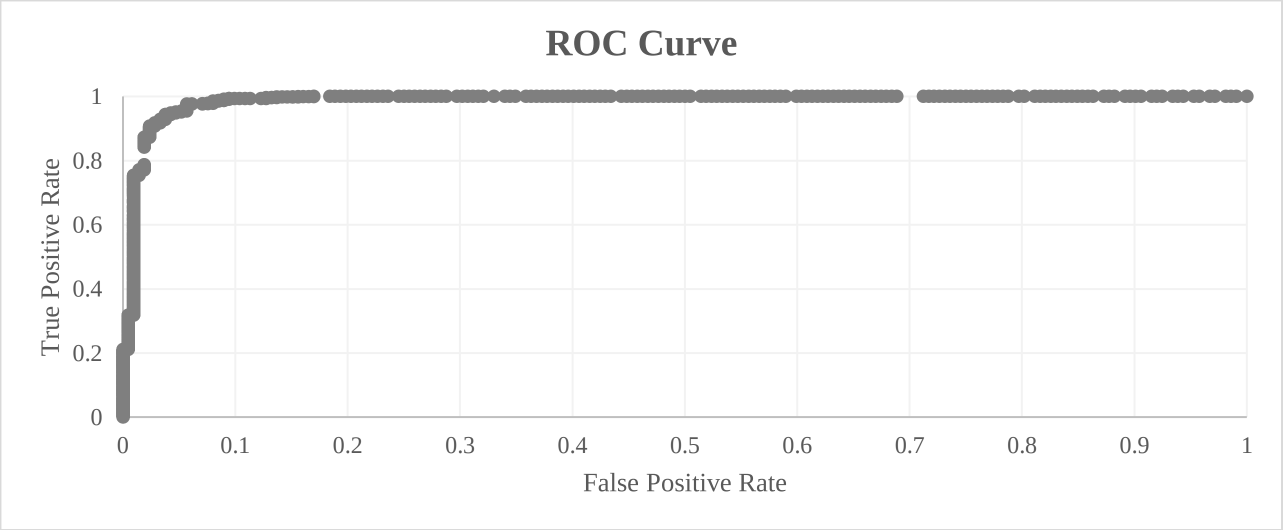
Confusion Matrix

- Accuracy = 98.15%
- Sensitivity = 99.38%
- Specificity = 89.62%

- The performance statistics of the resultant model exhibited an accuracy of 98.15% in detecting spam messages
- The sensitivity (Ham-detection capability) of the model was 99.38% with specificity (Spam-detection capability) of 89.62%

Appendix

The Naïve Bayes classifier trained on the obtained data and further tested revealed encouraging performance statistics



- The AUC value of the classification model was observed to be 0.9871 indicating high distinguishing capability between the classes

R Code & Dataset link

R Code: Spam Detection

Dataset : SMS Spam Collection Dataset

References

- Is Naïve Bayes a Good Classifier for Document Classification?

  S.L. Ting, W.H. Ip, Albert H.C. Tsang

- Bayes' Theorem & Naïve Bayes Classifier

  Daniel Berrar

- Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking

  Chanju Kim and Kyu-Baek Hwang