AI VIETNAM
All-in-One Course
(TA Session)

Text Classification

Project

Dinh-Thang Duong – TA

Anh-Khoi Nguyen – STA

Year 2024

Getting Started

❖ Objectives

$$p(Y|X) = \frac{p(X|Y) \times p(Y)}{p(X)}$$

Spam	Not spam
	
"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6 days, 16+ TsandCs apply Reply HL 4 info"	"Nah I don't think he goes to usf, he lives around here though"

Our objectives:

- Discuss about Naïve Bayes algorithm.
- Delve into one of popular NLP tasks: Text Classification.
- Apply Naïve Bayes and its variants to solve a text classification task.
- Investigate an improved baseline over Naïve Bayes and conduct experiments.

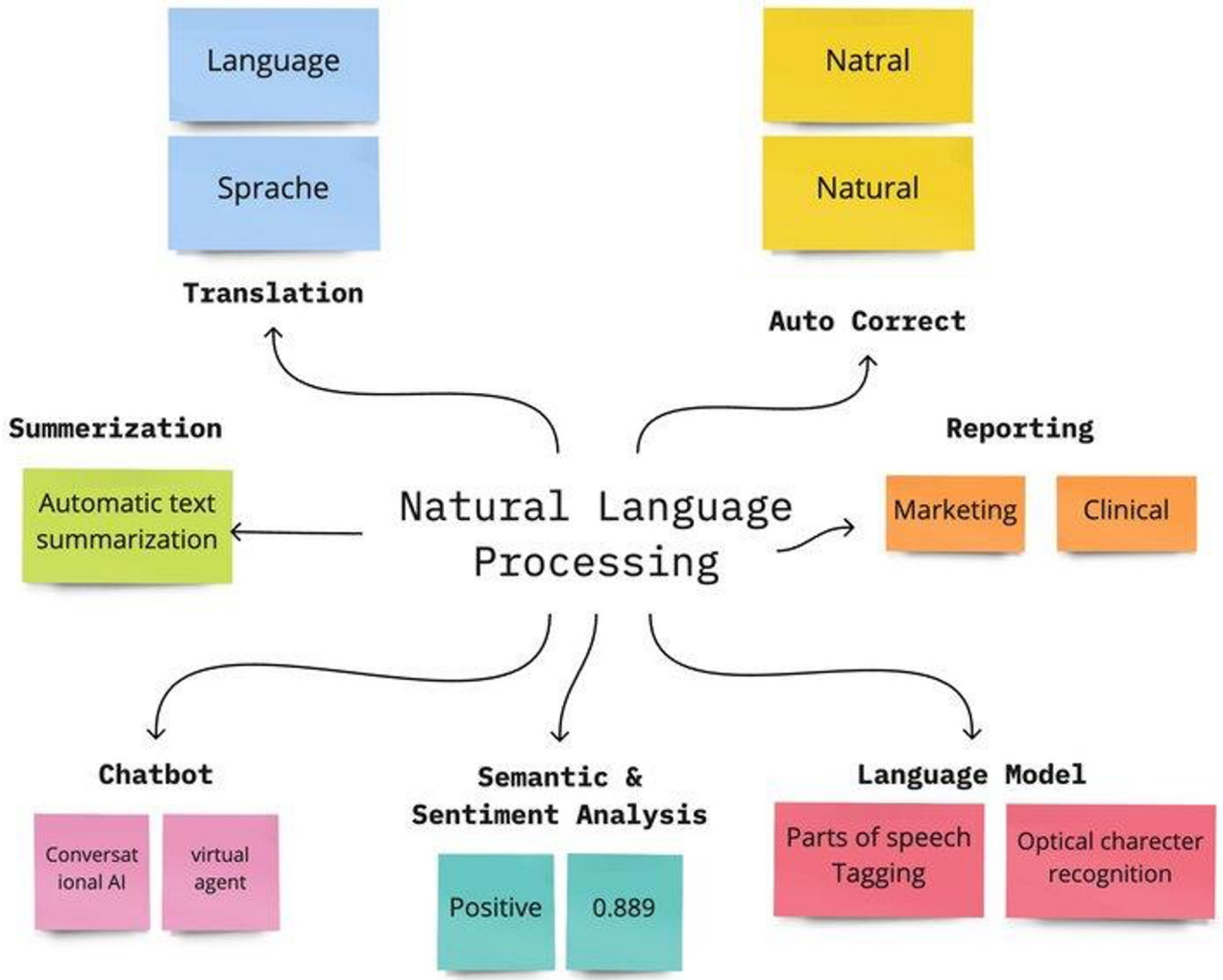
Outline

- Introduction
- Message Classification
- Code Implementation
- Improved Baseline
- Question

Introduction

Introduction

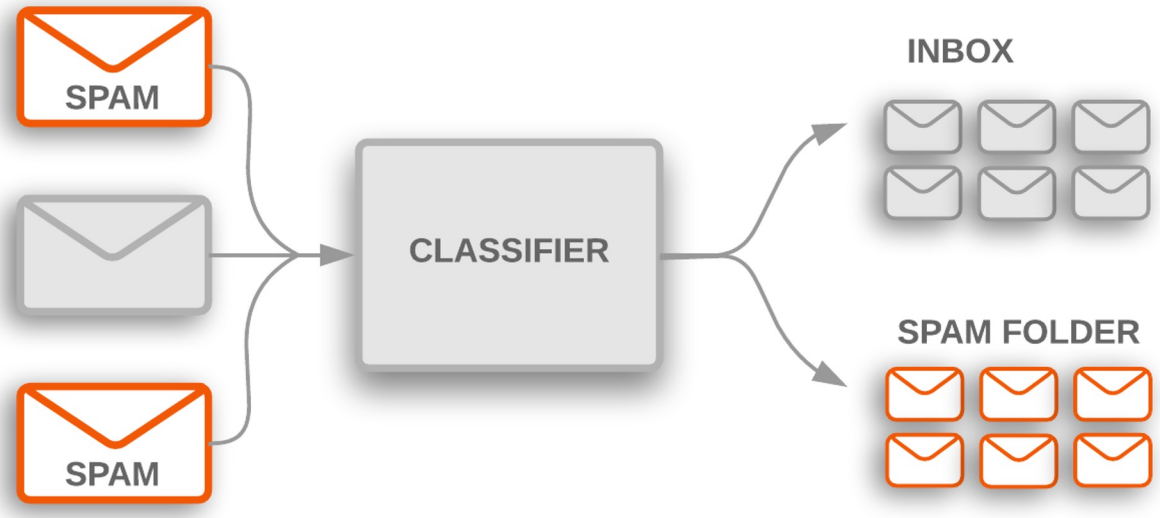
❖ Getting Started



https://www.researchgate.net/figure/Natural-Language-Processing_fig1_364842359

Introduction

❖ Text Classification

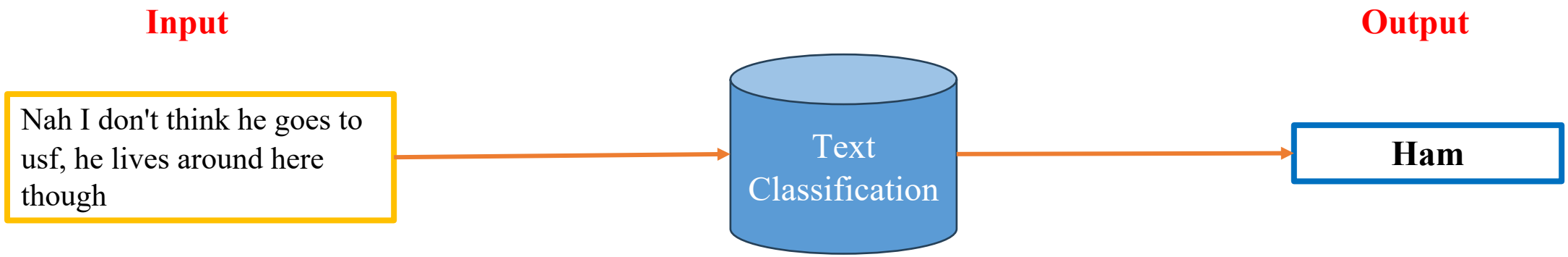


Text classification: A Natural Language Processing (NLP) task that involves categorizing text into predefined labels or classes. It is used to automatically assign a category to a text document, such as spam detection in emails, sentiment analysis of reviews, or topic classification of articles.

https://developers.google.com/machine-learning/guides/text-classification

Introduction

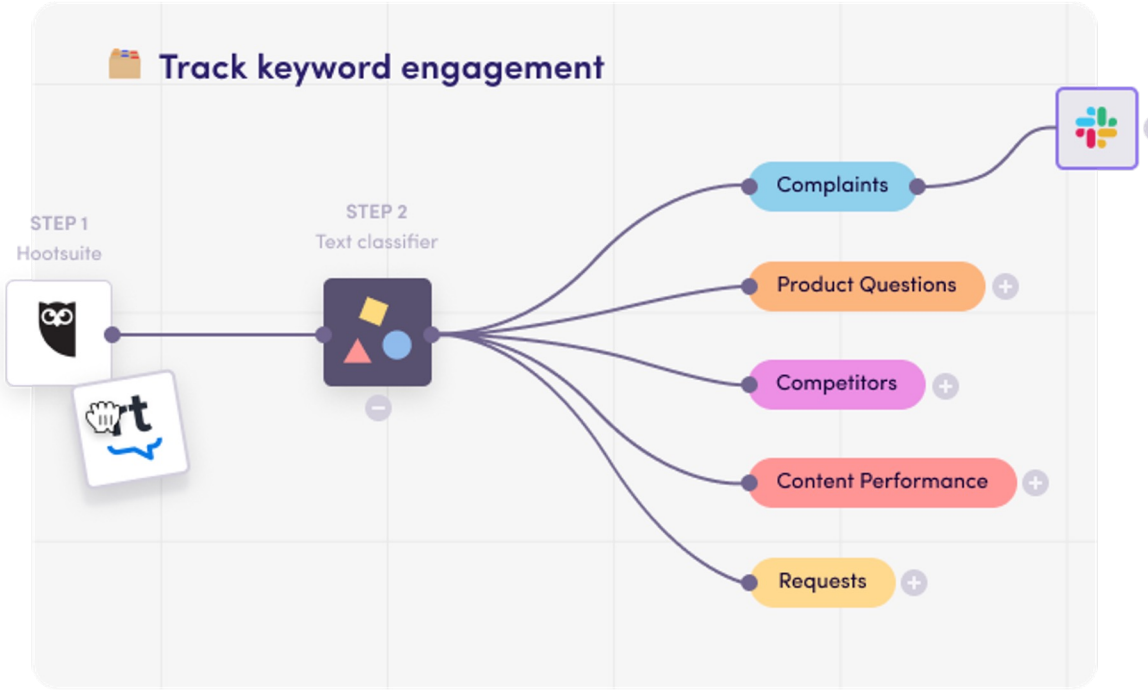
❖ Text Classification I/O



Text Classification: **An NLP task that aims to classify a given text into pre-defined classes.**

Introduction

❖ Applications

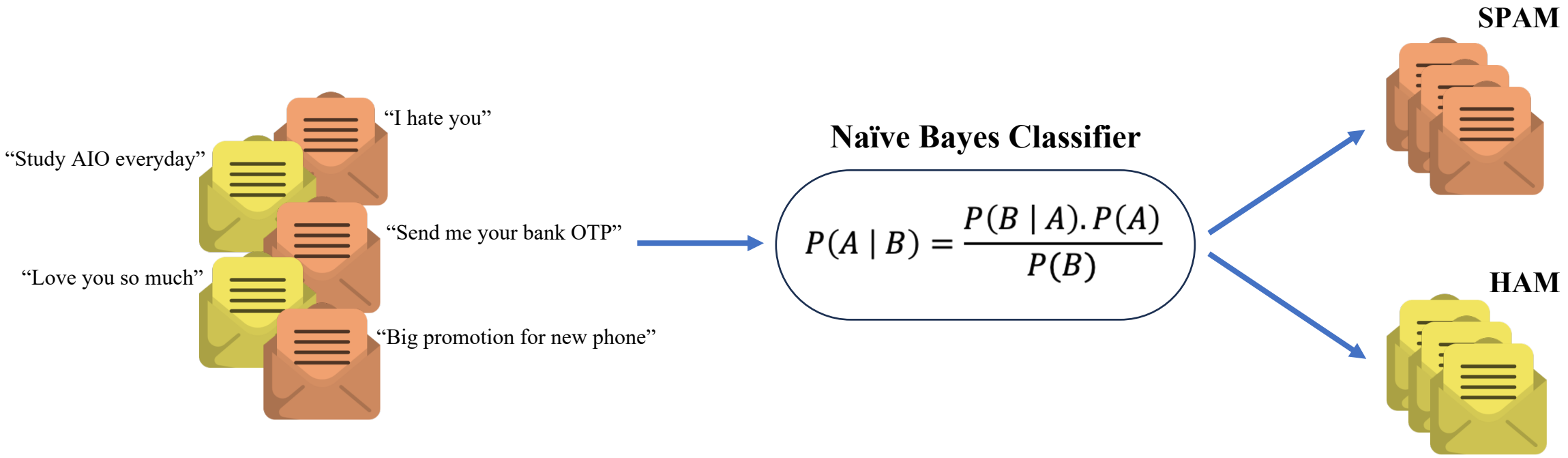


https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/
https://levity.ai/blog/9-text-classification-examples

Introduction

❖ Project Statement: Text Classification Naïve Bayes

Description: Given **Message Classification Dataset**, build a Naives Bayes model to determine whether a text message is spam message or not (ham).



https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

Introduction

❖ Message Classification I/O

Spam	Not spam
	
"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info"	"Nah I don't think he goes to usf, he lives around here though"

Problem Statement: Given a message, classify it into one of the two classes: Spam or Ham

Input: "Nah I don't think he goes to usf, he lives around here though"
Output: "Ham"

Message Classification Problem

Message Classification

Message Classification

❖ Probability

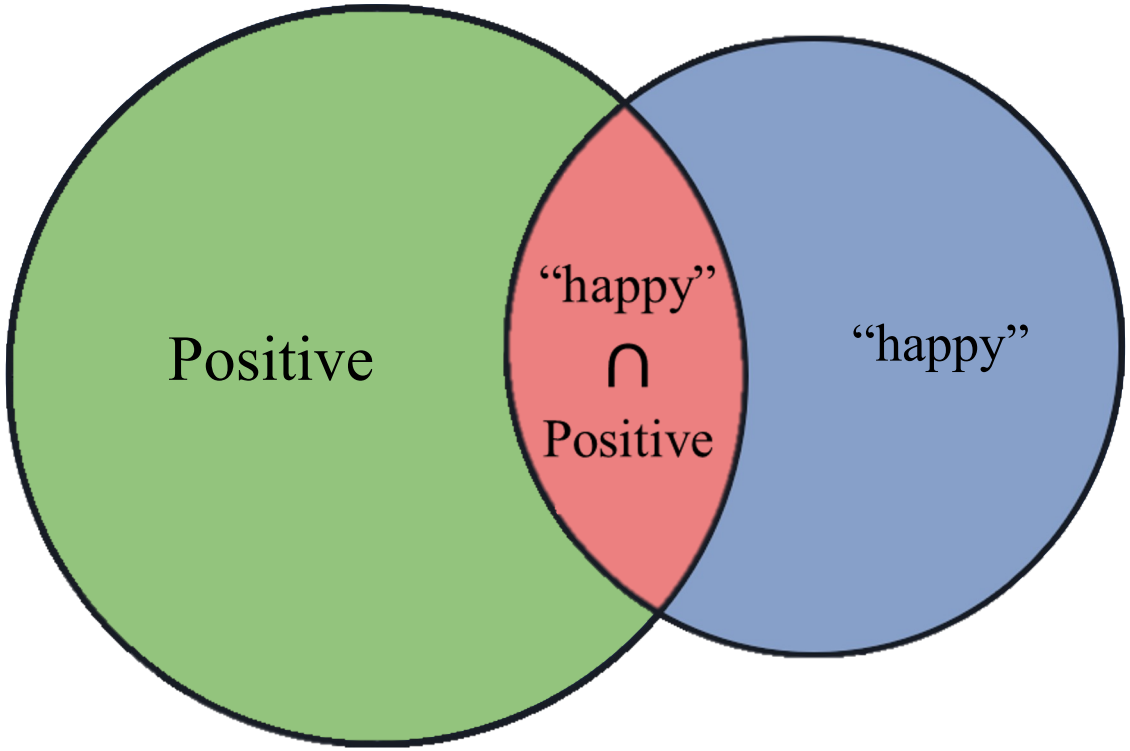
Positive	Positive	Positive	Positive	Positive
Positive	Positive	Positive	*Negative*	*Negative*
Negative	*Negative*	*Negative*	*Negative*	*Negative*
Negative	*Negative*	*Negative*	*Negative*	*Negative*

$$P(Pos) = \frac{N_{Pos}}{N} = \frac{8}{20} = 0.4$$

$$P(Neg) = 1 - P(Pos) = 1 - 0.4 = 0.6$$

Message Classification

❖ Conditional Probability



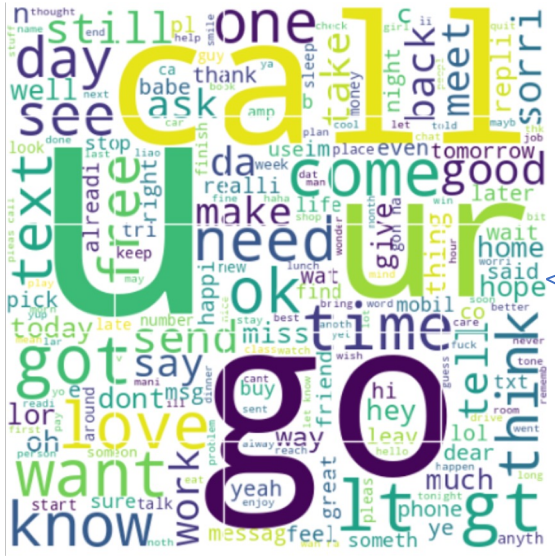
$$P("happy" \mid Pos) = \frac{P(Pos \cap "happy")}{P(Pos)}$$

Message Classification

❖ Bayes' Rule

The content of a message consists of many **words** combined together

How do we know which letter is SPAM or HAM?



SPAM

"hate", "promotion", "sale", ...

? **How do you know which group a word belongs to?**

HAM

"happy", "love", "deadline",...

One of the ways is based on individual words

Message Classification

❖ Bayes' Rule

How do you know which group a word belongs to?

Bayes' Rule: $$P(A \mid B) = \frac{P(B \mid A).P(A)}{P(B)}$$

$$P(Class \mid Words) = \frac{P(Words \mid Class)\ .P(Class)}{P(Words)}$$

$$P(Ham \mid w_1, \dots, w_n) = \frac{P(w_1, \dots, w_n \mid Ham).P(Ham)}{P(w_1, \dots, w_n)}$$

!!!

$$P(Ham \mid \text{"happy"}, \text{"hate"}, \text{"nice"}, \dots) = \frac{\mathbf{P(\text{"happy"}, \text{"hate"}, \text{"nice"}, \dots \mid Ham)}.P(Ham)}{P(\text{"happy"}, \text{"hate"}, \text{"nice"}, \dots)}$$

Message Classification

❖ Word Dependence

$$P("brown","sugar","pearl","milk","tea"|\ Ham)$$

Human:



"brown sugar pearl milk tea"

Naïve:





$$P("brown"|\ Ham).P("sugar"|\ Ham).P("pearl"|\ Ham).P("milk"|\ Ham).P("tea"|\ Ham)$$

Message Classification

❖ Naïve Bayes

Bayes' rule with the "Naïve" is the presence of one feature (word) does not affect the presence of another (word).

$$P("happy","hate","nice",\dots \mid Ham) = \prod_{i=1}^{n} P(word_i \mid y) =$$

$$P("happy"|Ham).P("hate"|Ham).P("nice"|Ham).P(" \dots "|Ham)$$

$$P(Ham \mid "happy","hate","nice",\dots) = \frac{\prod_{i=1}^{n} \boldsymbol{P(word_i \mid Ham)} \,.P(Ham)}{P("happy","hate","nice",\dots)}$$

Message Classification

❖ Remove constant

This value does not change when we consider different y classes.

$$P(Ham \mid "happy", "hate", "nice", \dots) = \frac{\prod_{i=1}^{n} P(word_i \mid Ham) \,.\, P(Ham)}{\boldsymbol{P("happy", "hate", "nice", \dots)}}$$

$$P(Spam \mid "happy", "hate", "nice", \dots) = \frac{\prod_{i=1}^{n} P(word_i \mid Spam) \,.\, P(Spam)}{\boldsymbol{P("happy", "hate", "nice", \dots)}}$$

Message Classification

❖ Remove constant

This value does not change when we consider different y classes.

$$P(Ham \mid "happy", "hate", "nice", \dots) = \frac{\prod_{i=1}^{n} P(word_i \mid Ham) \,.\, P(Ham)}{P("happy", "hate", "nice", \dots)}$$

$$P(Spam \mid "happy", "hate", "nice", \dots) = \frac{\prod_{i=1}^{n} P(word_i \mid Spam) \,.\, P(Spam)}{P("happy", "hate", "nice", \dots)}$$

Naïve Bayes Message Classification :

$$P(y \mid \boldsymbol{words}) \propto P(y) \,.\, \prod_{i=1}^{n} P(word_i \mid y)$$

Message Classification

❖ Project Statement

Description: Given **Message Classification Dataset**, build a Naives Bayes model to determine whether a text message is spam message or not (ham).

SPAM	HAM (NOT SPAM)
	
"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info"	"Nah I don't think he goes to usf, he lives around here though"

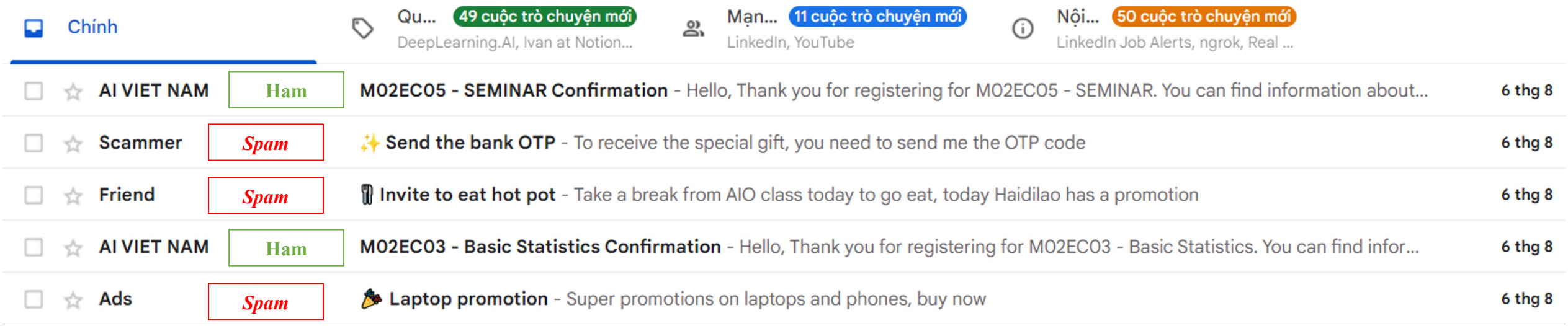
Message Classification

❖ Introduction

Email classification is a task where we determine whether a given Message is

- Spam (unsolicited or unwanted) *Spam*
- Ham (non-spam) Ham

Suppose you are "all in" AIO and you receive the following emails:



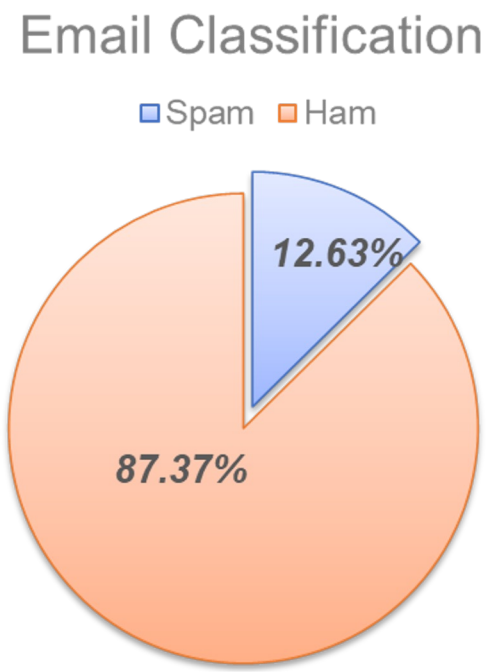
Message Classification

❖ Dataset

The data set includes 2 columns:

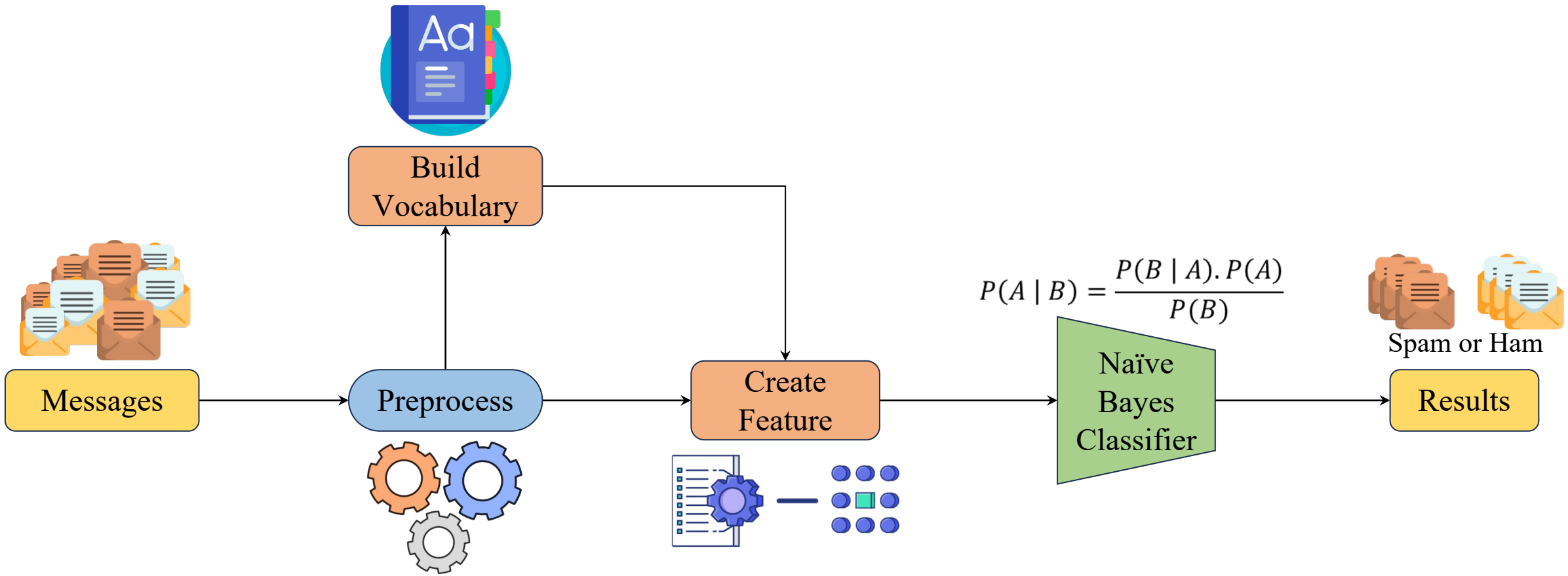
- **Category** (mail type)
- **Message** (mail content)

Category	Message
ham 87% spam 13%	5157 unique values
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for



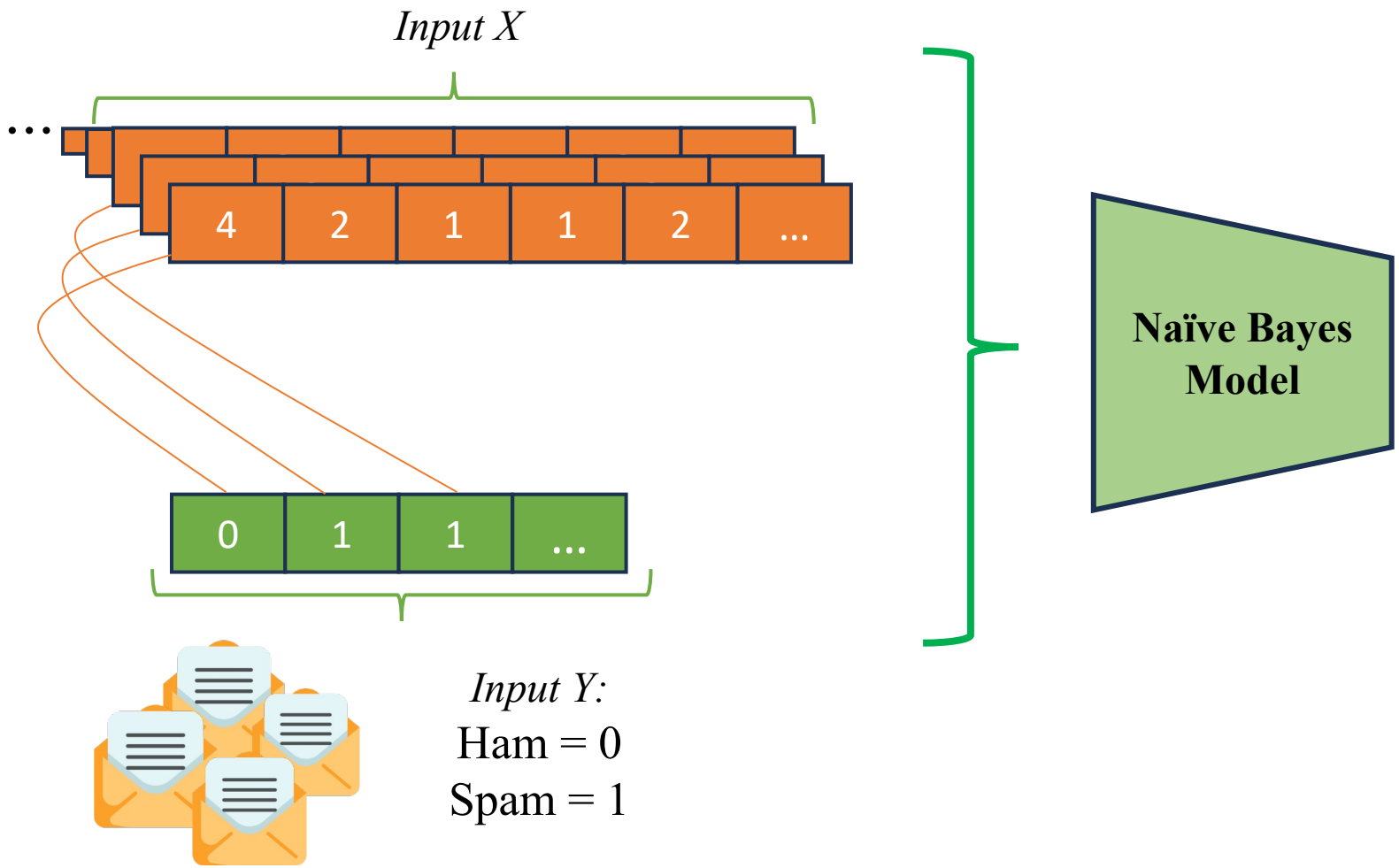
Message Classification

❖ Project Pipeline



Message Classification

❖ Input Naïve Bayes Model

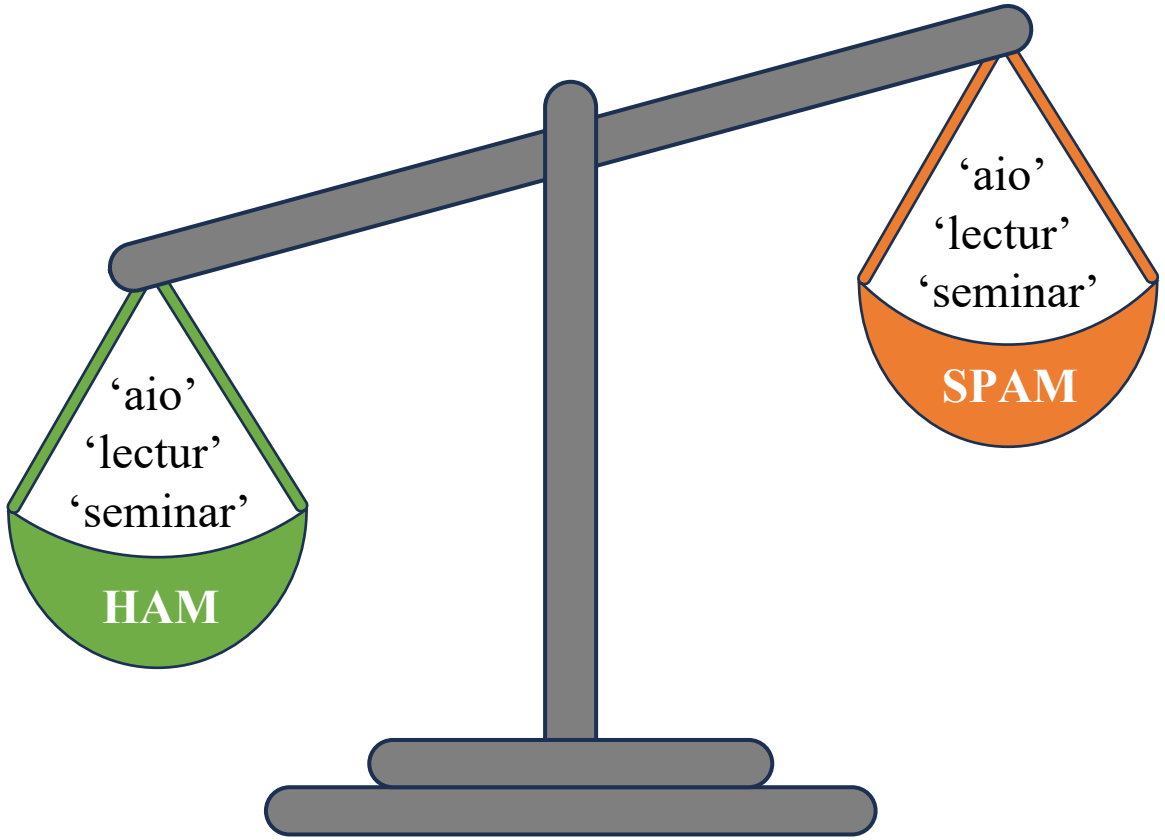


Message Classification

❖ Applying Naïve Bayes with Maximum A Posteriori Estimation

To classification, using MAP to estimation $P(y)$ and $P(word_i \mid y)$, where y is a label in training set.
Which y has the highest probability is the model result.

$$\hat{y} = \underset{y}{arg\,max}\, P(y) . \prod_{i=1}^{n} P(word_i \mid y)$$



Message Classification

❖ $P(x_i|y)$

$$\hat{y} = arg\max_{y} P(y) . \prod_{i=1}^{n} P(x_i \mid y)$$

Suppose Input X:

Suppose Input Y:

Word	Ham	Spam
aio	4 + 0 = 4	0
i	2 + 1 = 3	3
studi	1 + 1 = 2	0
we	1 + 0 = 1	2
love	2 + 1 = 3	0
V	**4+3+2+1+3=13**	**3+2=5**

Message Classification

❖ $P(x_i|y)$

Naïve Bayes Model

$$\hat{y} = arg\max_y P(y) . \prod_{i=1}^{n} P(x_i \mid y)$$

Word	Ham	Spam
aio	4 / 13	0 / 5
i	3 / 13	3 / 5
studi	2 / 13	0 / 5
we	1 / 13	2 / 5
love	3 / 13	0 / 5
V	**13**	**5**

Word	Ham	Spam
aio	0.308	0
i	0.231	0.6
studi	0.154	0
we	0.077	0.4
love	0.231	0
V	**13**	**5**

Message Classification

❖ $P(y)$

Naïve Bayes Model

$$\hat{y} = arg\max_{y} P(y) . \prod_{i=1}^{n} P(x_i \mid y)$$

Suppose Input Y:

0	0	1

$$P(Ham) = \frac{2}{N} = \frac{2}{3} = 0.667$$

$$P(Spam) = \frac{1}{N} = \frac{1}{3} = 0.334$$

Message Classification

❖ Predict

Naïve Bayes Model

Word	Ham	Spam
aio	0.308	0
i	0.231	0.6
studi	0.154	0
we	0.077	0.4
love	0.231	0
V	**13**	**5**

$P(Ham) = 0.667$

$P(Spam) = 0.334$

Predict

New message: "**I study AIO, I love it**"

$$\hat{y} = arg\max_{y} P(y) . \prod_{i=1}^{n} P(word_i \mid y)$$

- $\hat{y}_{ham} = (0.667) * (0.231 * 0.154 * 0.308 * 0.231 * 0.231) \approx 0.00038997$
- $\hat{y}_{spam} = (0.334) * (0.6 * 0.1 * 0.1 * 0.6 * 0.1) \approx 0.00012$

$$\hat{y}_{ham} > \hat{y}_{spam} \rightarrow \text{Predict} = \text{Ham}$$

Message Classification

❖ Problem 1 – Zero Probability

Naïve Bayes Model

$$\hat{y} = arg\max_{y} P(y) \cdot \prod_{i=1}^{n} P(word_i \mid y)$$

Word	Ham	Spam
aio	4 / 13	0 / 5
i	3 / 13	3 / 5
studi	2 / 13	0 / 5
we	1 / 13	2 / 5
love	3 / 13	0 / 5
V	13	5

Word	Ham	Spam
aio	0.308	0 !
i	0.231	0.6
studi	0.154	0 !
we	0.077	0.4
love	0.231	0 !
V	13	5

Message Classification

❖ Problem 1 - Laplacian Smoothing

Word	Ham	Spam
aio	4	0
i	3	3
studi	2	0
we	1	2
love	3	0
V	13	5

$$P(x_i \mid y) = \frac{P(x_i) + 1}{V + N_{dictionary}}$$

Word	Ham	Spam
aio	(4+1)/(13+5)	(0+1)/(5+5)
i	(3+1)/(13+5)	(3+1)/(5+5)
studi	(2+1)/(13+5)	(0+1)/(5+5)
we	(1+1)/(13+5)	(2+1)/(5+5)
love	(3+1)/(13+5)	(0+1)/(5+5)
V	13	5

Message Classification

❖ Problem 1 - Laplacian Smoothing

$$P(x_i \mid y) = \frac{P(x_i) + 1}{V + N_{dictionary}}$$

Word	Ham	Spam
aio	(4+1)/(13+5)	(0+1)/(5+5)
i	(3+1)/(13+5)	(3+1)/(5+5)
studi	(2+1)/(13+5)	(0+1)/(5+5)
we	(1+1)/(13+5)	(2+1)/(5+5)
love	(3+1)/(13+5)	(0+1)/(5+5)

Word	Ham	Spam
aio	0.278	0.1
i	0.222	0.4
studi	0.167	0.1
we	0.111	0.3
love	0.222	0.1
Sum	1	1

Message Classification

❖ Problem 1 - Laplacian Smoothing

Word	Ham	Spam
aio	0.278	0.1
i	0.222	0.4
studi	0.167	0.1
we	0.111	0.3
love	0.222	0.1
Sum	1	1

Predict

New message: "**I study AIO, I love it.**"

- $\hat{y}_{ham} = (0.667) * (0.222 * 0.167 * 0.278 * 0.222 * 0.222) \approx 0.0003388$
- $\hat{y}_{spam} = (0.334) * (0.4 * 0.1 * 0.1 * 0.4 * 0.1) \approx 0.00005344$

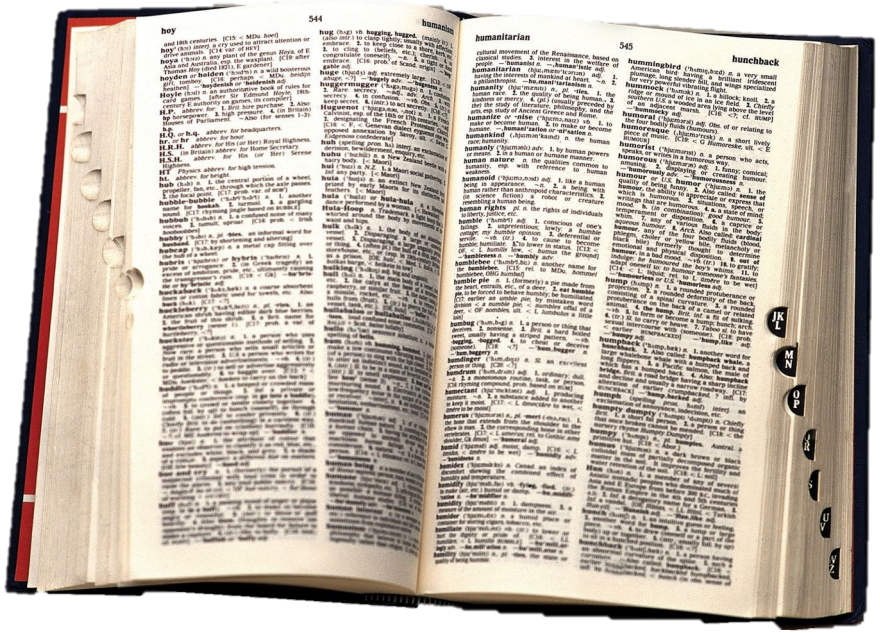
$\hat{y}_{ham} > \hat{y}_{spam} \rightarrow$ Predict = Ham

Message Classification

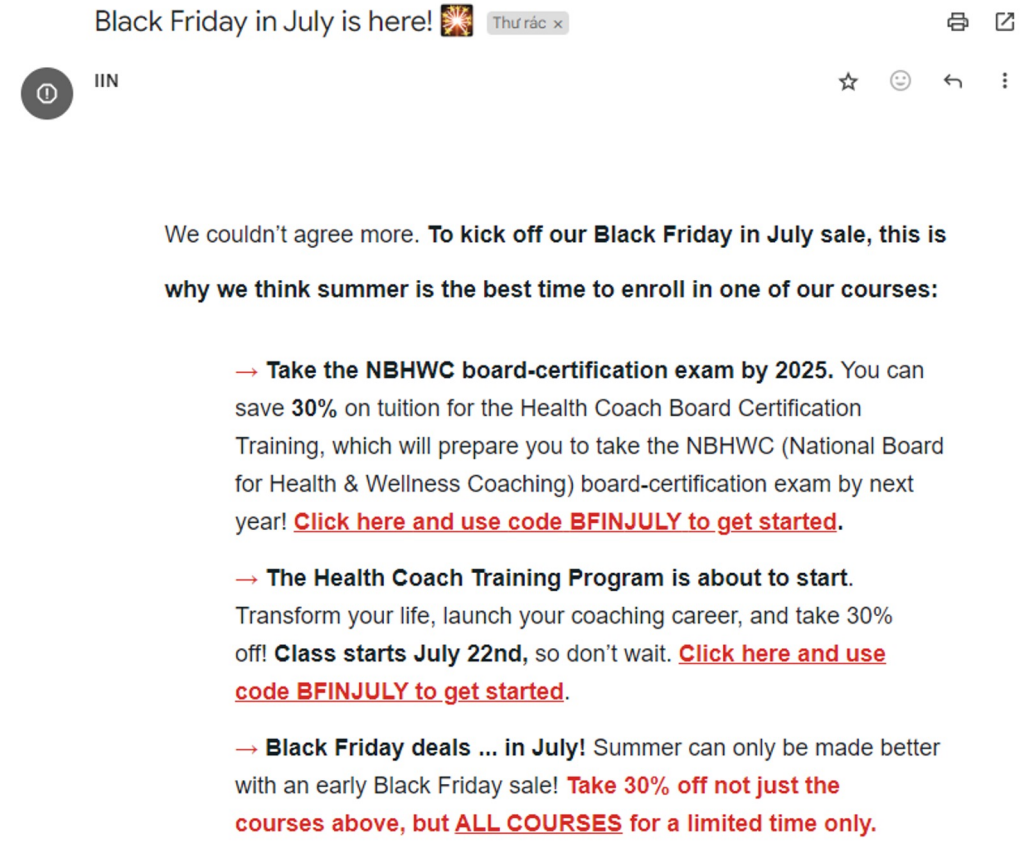
❖ Problem 2 – Log Likelihood

$\hat{y}_{spam} \approx 0.00005344$! $\hat{y}_{ham} \approx 0.0003388$!

In reality, the dictionary is huge



and the message content much longer



The results will become very very small, gradually approaching 0

Message Classification

❖ Problem 2 – Log Likelihood

Take advantage of the **Scaling** property of *Logarithms*.

x	$Log_{10}(x)$
0.00001	-5
0.001	-3
0	Error
100	3
10000	5

$$\hat{y} = arg\max_y P(y) \prod_{i=1}^{n} P(w_i|y)$$

$$= arg\max_y \log P(y) \prod_{i=1}^{n} P(w_i|y)$$

$$= arg\max_y \log P(y) + \log \prod_{i=1}^{n} P(w_i|y)$$

$$= arg\max_y \log P(y) + \log \sum_{i=1}^{n} P(w_i|y)$$

since $\log(ab) = \log(a) + \log(b)$

If a < b, log(a) < log(b)
$\Rightarrow$ change from prob to log-prob.
$\Rightarrow$ $(0, 1)$ to $(-\infty, 0)$

Message Classification

❖ Problem 2 – Log Likelihood

Word	Ham	Spam
aio	0.278	0.1
i	0.222	0.4
studi	0.167	0.1
we	0.111	0.3
love	0.222	0.1

Predict

New message: "**I study AIO, I love it.**"

$$\hat{y}_{ham} = [\log(0.222) + \log(0.167) + \log(0.278) + \log(0.222) + \log(0.222)] + \log(0.667)$$

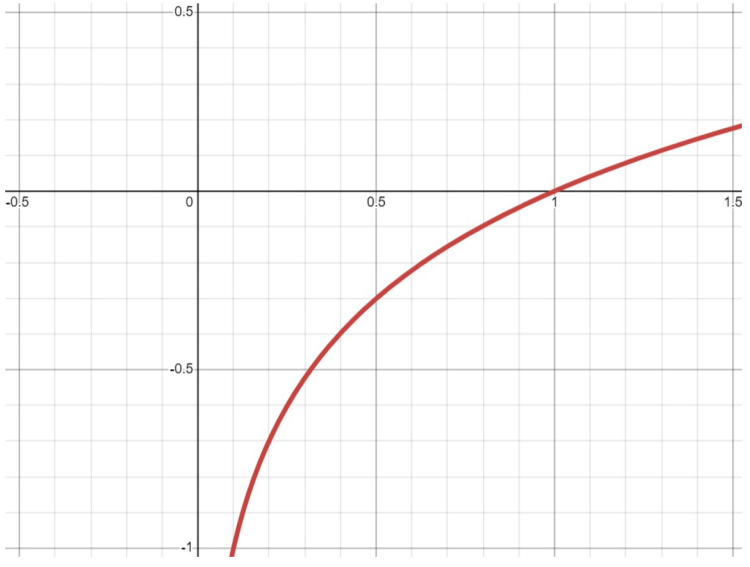
$$\approx -3.294 - 0.175 \approx -3.469$$

$$\hat{y}_{spam} = [\log(0.4) + \log(0.1) + \log(0.1) + \log(0.4) + \log(0.1)] + \log(0.334)$$

$$\approx -3.796 - 0.476 \approx -4.272$$

Since a < b, log(a) < log(b) => A smaller number (more negative) corresponds to a smaller probability, hence:

$$\hat{y}_{ham} > \hat{y}_{spam} \rightarrow \text{Predict} = \text{Ham}$$



Break - Quiz

Break - Quiz

Câu 1: Mục tiêu chính của bài toán Text Classification là gì?
A. Tóm tắt nội dung văn bản.
B. Dịch văn bản sang ngôn ngữ khác.
C. Phân loại văn bản vào các nhóm định trước.
D. Tạo ra văn bản mới dựa trên văn bản đã cho.

Câu 2: Trong bài toán phân loại email, hai lớp chính thường được sử dụng là gì?
A. Quan trọng và Không quan trọng.
B. Spam và Ham (không phải spam).
C. Cá nhân và Công việc.
D. Đã đọc và Chưa đọc.

Câu 3: Giả định "Naive" trong thuật toán Naive Bayes là gì?
A. Tất cả email đều có xác suất như nhau để là Spam.
B. Độ dài của email quyết định nó là Spam hay Ham.
C. Sự xuất hiện của một từ không ảnh hưởng đến sự xuất hiện của các từ khác.
D. Các email Spam luôn chứa các từ khóa cụ thể.

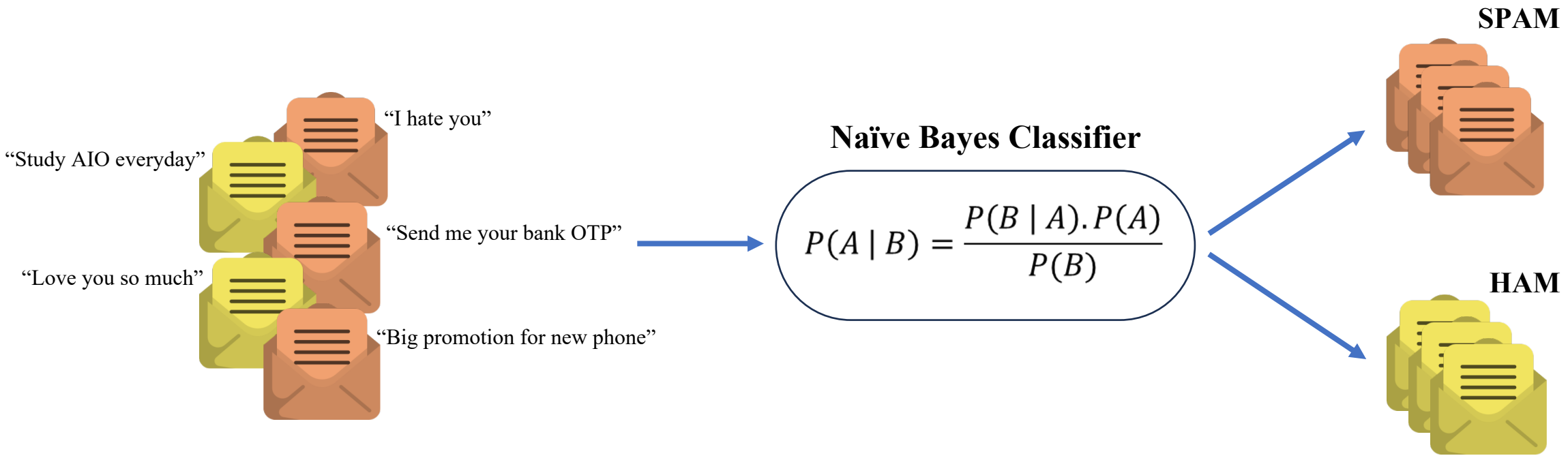
Câu 4: Khi áp dụng Naive Bayes cho phân loại email, tại sao có thể bỏ qua mẫu số P(w) trong công thức Bayes?
A. Vì P(w) luôn bằng 1.
B. Vì P(w) không ảnh hưởng đến việc so sánh xác suất giữa các lớp.
C. Vì P(w) chỉ quan trọng trong trường hợp có nhiều hơn 2 lớp.
D. Vì P(w) luôn nhỏ hơn P(c|w).

Code Implementation

Code Implementation

❖ Introduction

Description: Given **Message Classification Dataset**, build a Naives Bayes model to determine whether a text message is spam message or not (ham).



https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf