A Maximum Entropy Model for Text Classification

Nguyen Viet Cuong
Nguyen Thi Thuy Linh
Ha Quang Thuy
Phan Xuan Hieu

Speaker:葉昱廷

SLIDE 2

ABSTRACT

This paper presents a machine learning model for Vietnamese text/web content classification and filtering that is based on the maximum entropy principle

The difficulty in identifying word boundaries of Vietnamese (isolated language) is solved by Maximum Matching approach based on a Vietnamese lexicon (LacViet MTD)

The Power of PowerPoint

Vietnamese Example

Vietnamese is an isolated language and whitespaces are not always used to identify the word boundaries

How are you doing?

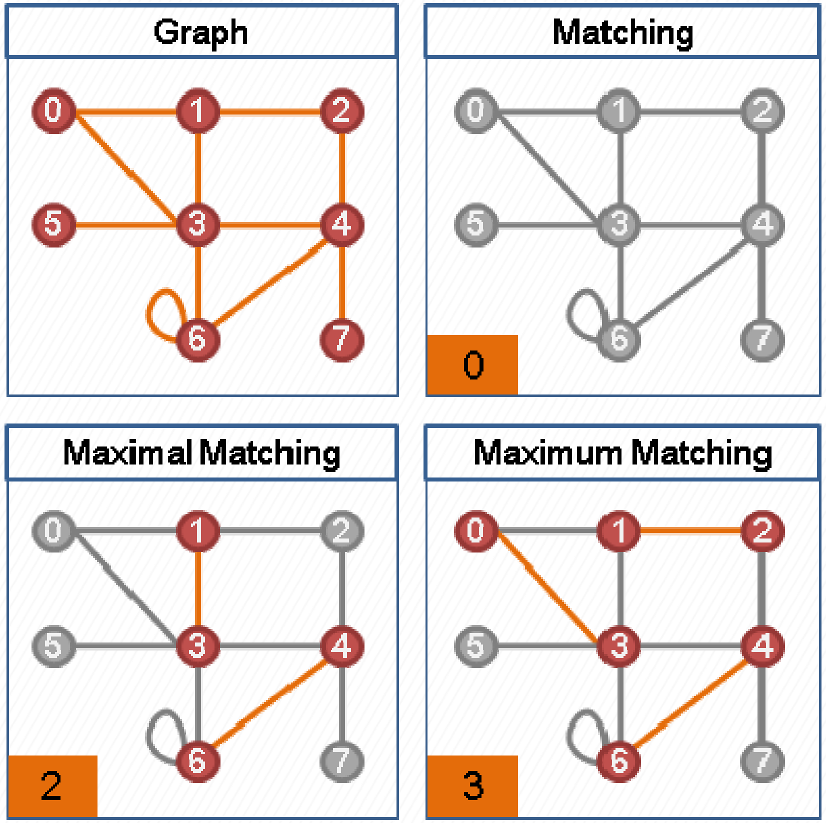
Dạo này chị thế nào ?

How # are # you # doing ?

Dạo # này chị # thế nào ?

SLIDE 4

Maximum Matching



Cardinality

Maximal matching：
沒有辦法直接增加配對數的匹配。

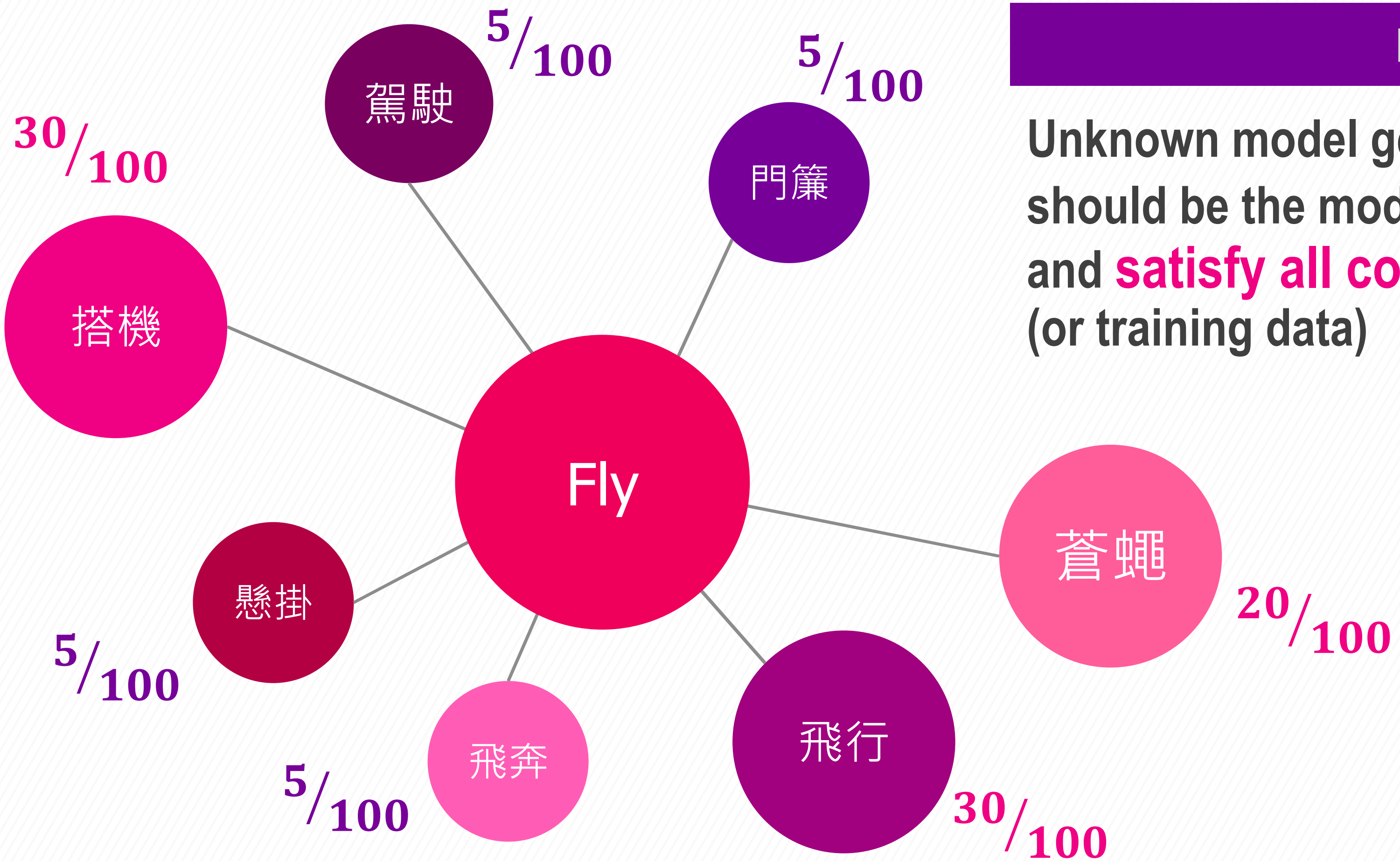
Maximum matching：
配對數最多的匹配。

Perfect matching：
所有點都配對到的匹配。

The Power of PowerPoint

SLIDE 5

Maximum Entropy



Fly

駕駛 5/100

門簾 5/100

搭機 30/100

懸掛 5/100

飛奔 5/100

飛行 30/100

蒼蠅 20/100

Main idea

Unknown model generating the sample data should be the model that is **most uniform** and **satisfy all constrains** from sample data (or training data)

The Power of PowerPoint

SLIDE 6

Maximum Entropy

Constrain Equation

$$\sum_{d,c}\tilde{p}(d)p(c|d)f_i(d,c)=\sum_{d,c}\tilde{p}(d,c)f_i(d,c)$$

$D = \{(d_1,c_1), (d_2,c_2),\cdots, (d_N,c_N)\}$

d_i is list of *context predicate*

c_i is class corresponding to d_i

Exponential Form

$$p(c|d)=\frac{1}{Z(d)}\exp\left(\sum_i\lambda_i f_i(d,c)\right)$$

Normalization Factor

$$Z(d)=\sum_c\exp\left(\sum_i\lambda_i f_i(d,c)\right)$$

The Power of PowerPoint

Maximum Entropy

Entropy

$$H(p) = -\sum_{d,c} \tilde{p}(d)\, p(c \mid d) \log p(c \mid d)$$

Likelihood

$$L(p) = \sum_{d,c} \tilde{p}(d,c) \log p(d \mid c)$$

The same solution

$$p^{*} = \operatorname*{argmax}_{p} H(p) = \operatorname*{argmax}_{p} L(p)$$

Exponential form guarantees
that the likelihood surface is convex

N-Gram

N-gram of syllables

According to syllable , may contain meaningless words.

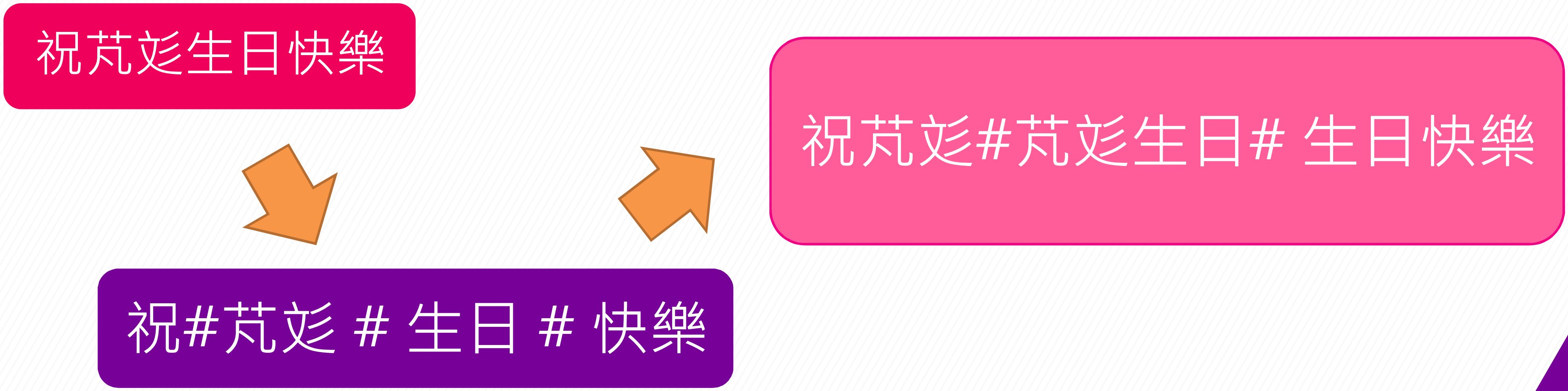
祝芃彣生日快樂

祝芃#芃彣 # 彣生 # 生日 # 日快 # 快樂

N-Gram

N-gram of words

According to segmentation , combination of meaningful words



祝芃彣生日快樂

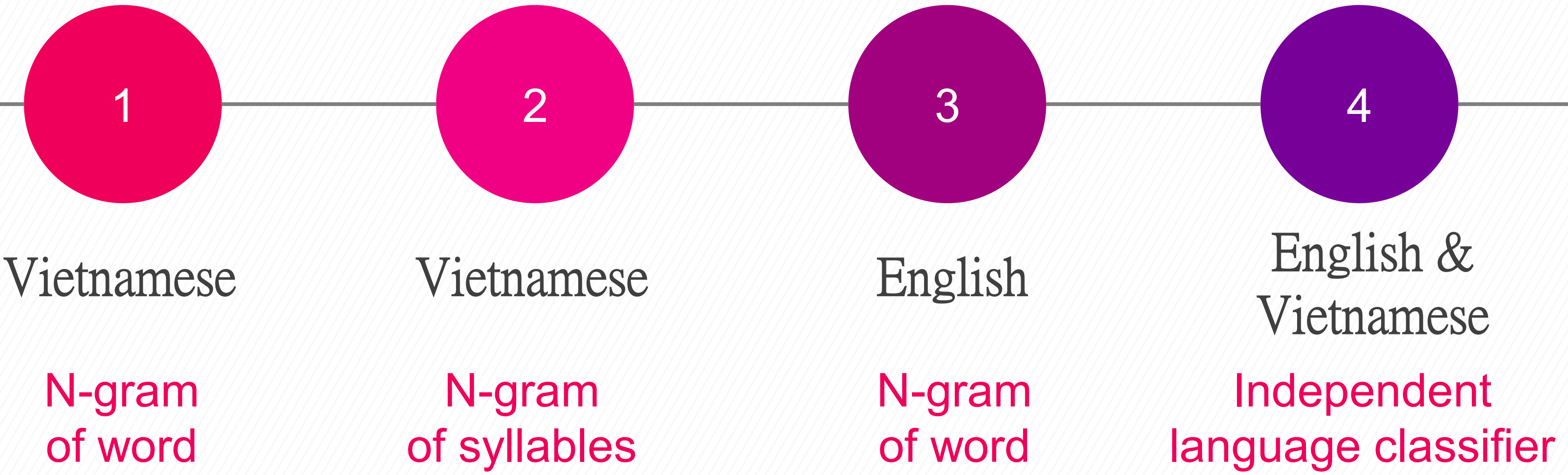
祝#芃彣 # 生日 # 快樂

祝芃彣#芃彣生日# 生日快樂

SECTION 03

Text Classification with ME

Experimental Setup



1

Vietnamese

N-gram of word

2

Vietnamese

N-gram of syllables

3

English

N-gram of word

4

English & Vietnamese

Independent language classifier

"10-fold cross validation test"

SLIDE 12

Training data

Vietnamese

6,400 **Vietnamese news pieces** in **8 classes**, so each class has 800 news pieces.

Vietnamese lexicon (LacViet MTD) with more than 70,000 entries

English

6,207 English news pieces in 8 classes

were collected from **BBC** News

The Power of PowerPoint

SLIDE 13

Building a Standard Class Tree

Vietnam Express News

Vietnam Net

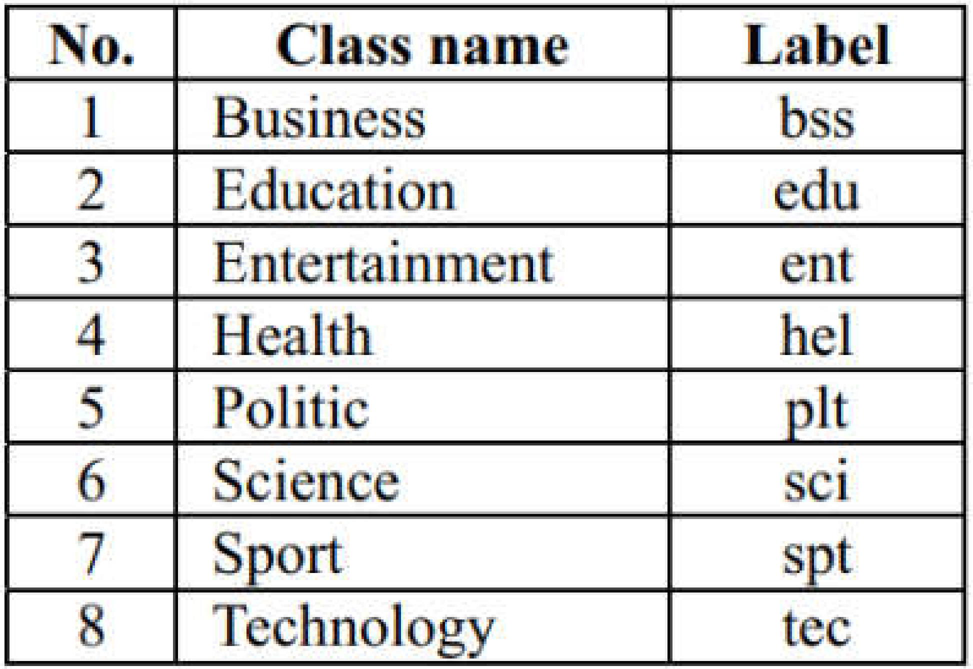
BBC News

CNN

The Power of PowerPoint

SLIDE 14

Class tree for text classification



No.	Class name	Label
1	Business	bss
2	Education	edu
3	Entertainment	ent
4	Health	hel
5	Politic	plt
6	Science	sci
7	Sport	spt
8	Technology	tec

For both Vietnamese and English training data.

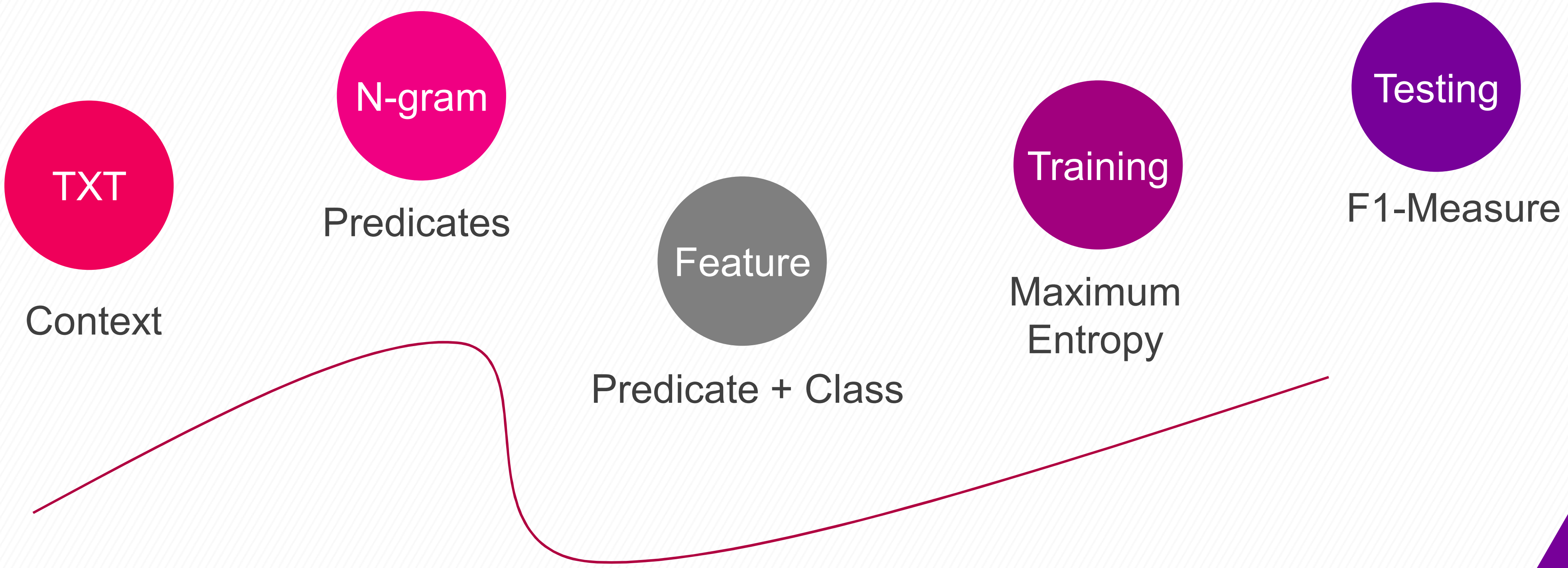
Feature Selection

Predicate	+	Class	=	Feature
công_nghệ		*tec*		*công_nghệ, tec*
công_nghệ		*edu*		*công_nghệ, edu*

Predicates come from N-gram of syllables or words

Features over predicates is approximate 5/3

Classification Procedures



TXT

Context

N-gram

Predicates

Feature

Predicate + Class

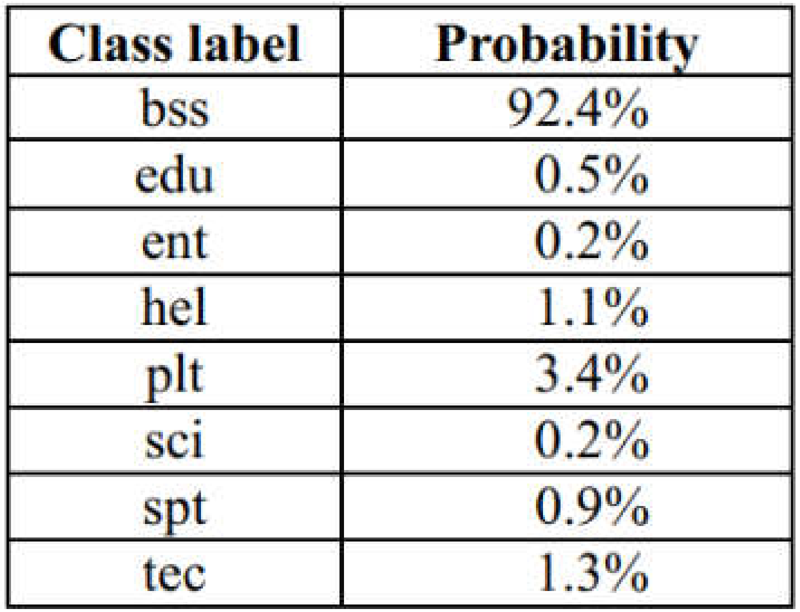
Training

Maximum Entropy

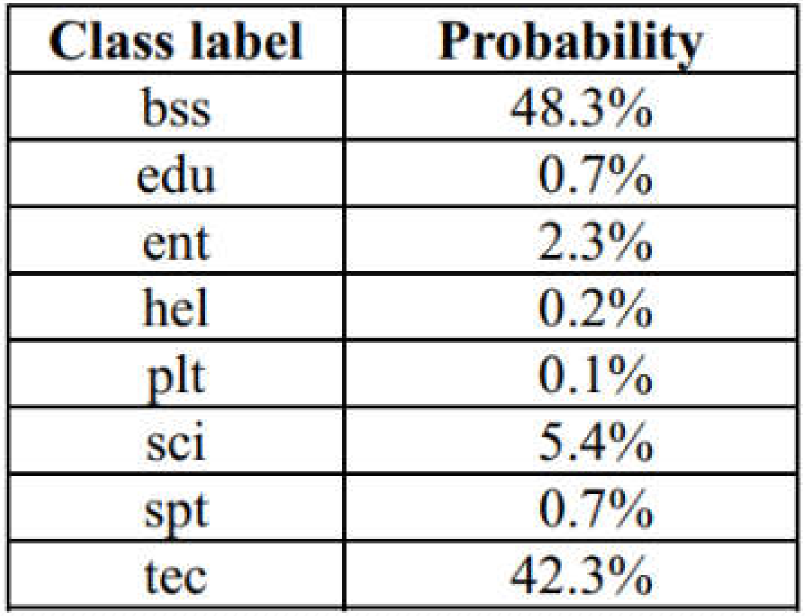
Testing

F1-Measure

Examples of probability distribution



Class label	Probability
bss	92.4%
edu	0.5%
ent	0.2%
hel	1.1%
plt	3.4%
sci	0.2%
spt	0.9%
tec	1.3%



Class label	Probability
bss	48.3%
edu	0.7%
ent	2.3%
hel	0.2%
plt	0.1%
sci	5.4%
spt	0.7%
tec	42.3%

Business

Business & Tech ?

Dynamic threshold

Threshold value is flexible and is different for each situation.

First

Choose p_m which is the highest probability value

Second

Calculate standard deviation

$$s_k = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(p_i - \bar{p})^2}$$

Last

Choose $p_j > p_m - S_k$

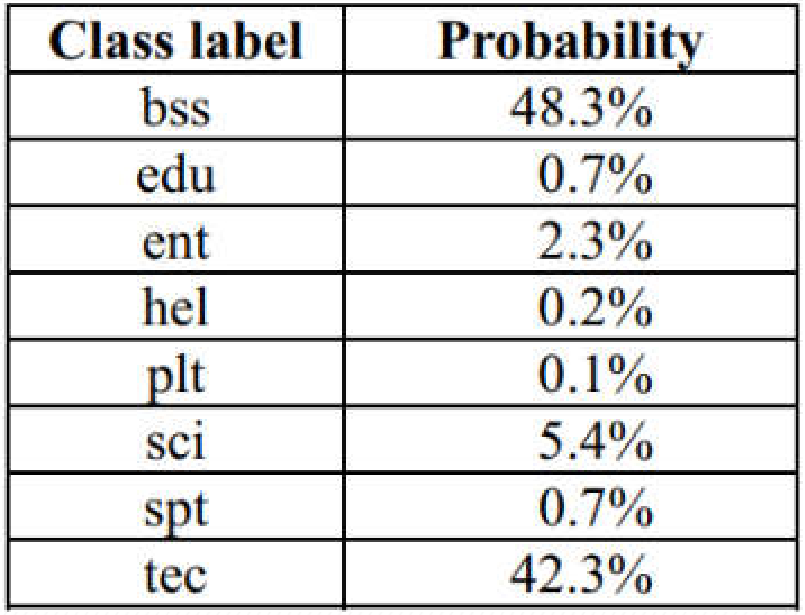
The corresponding class will be chosen

Firstly, we choose p_1= 48.3%

$$s_k = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(p_i - \overline{p})^2} \approx 19.1\%$$

Dynamic threshold t_d is :

$$t_d = p_1 - s_k = 48.3\% - 19.1\% = 29.2\%$$



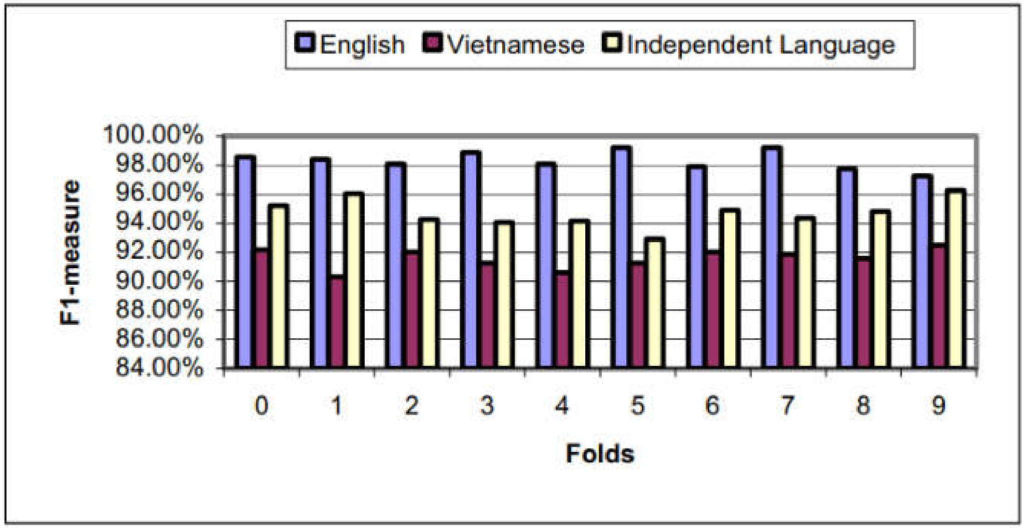
Class label	Probability
bss	48.3%
edu	0.7%
ent	2.3%
hel	0.2%
plt	0.1%
sci	5.4%
spt	0.7%
tec	42.3%

Business & Tech !

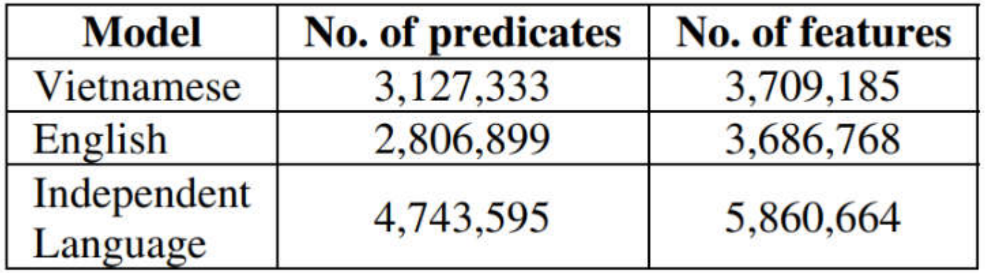
SLIDE 20

Results

F1 - Measure



No. of Predicates & Features



Model	No. of predicates	No. of features
Vietnamese	3,127,333	3,709,185
English	2,806,899	3,686,768
Independent Language	4,743,595	5,860,664

The ambiguity between English and Vietnamese is very low although both languages are written in Latin characters.

The Power of PowerP

That’s all. Thank you! ☺

Any Questions?