

# Đồ án cuối kỳ Data Science

Nhóm 12

Vũ Đăng Hoàng Long (18120203)

Nguyễn Huỳnh Đại Lợi (18120198)

# Nội dung

1. Ý tưởng
2. Quy trình làm việc
3. Khai thác
4. Khám phá
5. Tiền xử lý
6. Xây dựng và thử nghiệm mô hình
7. Tổng kết

Trước khi bắt đầu đồ án

# Ý tưởng

- Đề tài có thể làm? Nội dung có thể kiểm?


▷ Nhiều, đa dạng

- Nhưng khó tìm với ràng buộc parse HTML (dữ liệu lịch sử thường được xử lý sẵn cho tải về), API (\$\$\$)

Searching for predict within


 Comments 46,514

 Notebooks 19,048

 Topics 10,424

 Competitions 2,513

 Datasets 2,485

 Users 254

 Blogs 108

 Organizations 3

Filter by

#### Date

- ☐ Last 90 days 5,783
- ☐ Last week 432
- ☐ Today 61

#### Viewed By You

- ☐ Viewed 5
- ☐ Not Viewed 81,344

#### Dataset Size

- ☐ small 1,891
- ☐ medium 525
- ☐ large 67

#### Dataset File Types

- ☐ csv 1,907
- ☐ xlsx 155
- ☐ txt 104

[More](#)

#### Dataset License

- ☐ Other 1,621

81,349 Results

Sort by: Relevancy ▼



 Competition

### Predict Future Sales

Playground

in 2 years • 10129 teams

[Predict Future Sales](#)



 Competition

### Riiid! Answer Correctness Prediction

Featured • Kernels competition

8 days ago • \$100,000 • 3395 teams



 Competition

### Jane Street Market Prediction

Featured • Kernels competition

in a month • \$100,000 • 2459 teams



 Competition

### INGV - Volcanic Eruption Prediction

Playground

9 days ago • 620 teams

Get weather data for any location on the globe immediately with our superb **APIs**! Just **sign up** with your email and start using minute forecasts, hourly forecasts, history and other weather data in your applications for free. For more functionality, please consider our generous subscriptions.

Current weather and forecasts collection

Free	Startup	Developer	Professional	Enterprise
	40 USD / month	180 USD / month	470 USD / month	2,000 USD / month
Get API key	Subscribe	Subscribe	Subscribe	Subscribe
60 calls/minute 1,000,000 calls/month	600 calls/minute 10,000,000 calls/month	3,000 calls/minute 100,000,000 calls/month	30,000 calls/minute 1,000,000,000 calls/month	200,000 calls/minute 5,000,000,000 calls/month



# Ý tưởng

Dữ liệu dạng chữ rất nhiều

▷ Chủ đề liên quan tới NLP

▷ Phân loại văn bản là bài toán đơn giản phù hợp với nhóm

Input: đoạn văn bản bất kỳ

Output: chủ đề của văn bản

# Mục đích?

Khách quan: giúp phân loại chủ đề văn bản tự động, học từ nguồn dữ liệu khổng lồ trên mạng mà không cần phải có người giám sát

Chủ quan:

- Long: lọc tự động những bài viết mà em muốn trên Facebook để tránh lãng phí thời gian lướt Facebook 



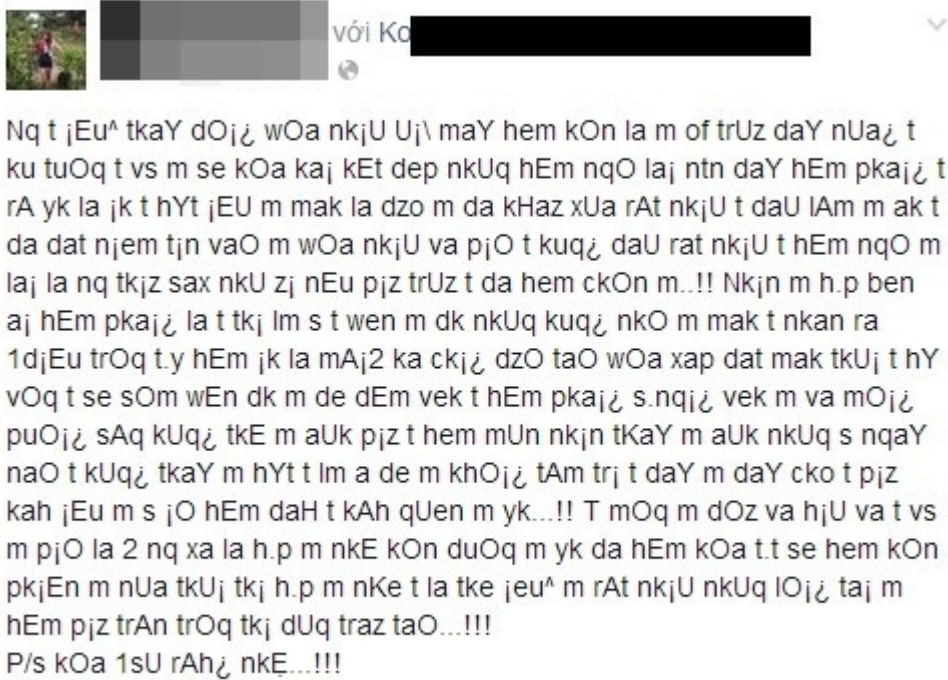
# Quy trình làm việc

- Nhóm bắt đầu muộn
- Dữ liệu lớn làm thời gian tại mỗi quy trình bị chậm.
  - ▷ Cần quy trình tối ưu
  - ▷ Mỗi thành viên đảm nhận một quy trình riêng và truyền thông lại kết quả cho nhau

Đồ án

# Khai thác dữ liệu

## Facebook?



A screenshot of a Facebook post. The post features a profile picture of a person in a field and a cover photo of a person in a field. The text of the post is in Vietnamese and appears to be a mix of random characters and words, possibly a placeholder or a test. The post is titled 'vời Kc' and has a dropdown arrow next to it. The text of the post is as follows:

Nq t jEu^ tkaY dOj\ wOa nkjU Uj\ maY hem kOn la m of trUz daY nUa\ t  
ku tuOq t vs m se kOa kaj kEt dep nkUq hEm nqO laj ntn daY hEm pkaj\ t  
rA yk la jk t hYt jEU m mak la dzo m da kHaz xUa rAt nkjU t daU lAm m ak t  
da dat njem tjn vaO m wOa nkjU va pjO t kuq\ daU rat nkjU t hEm nqO m  
laj la nq tkjz sax nkU zj nEu pjz trUz t da hem ckOn m...!! Nkin m h.p ben  
aj hEm pkaj\ la t tkj lm s t wen m dk nkUq kuq\ nkO m mak t nkan ra  
1djEu trOq t.y hEm jk la mAj2 ka ckj\ dzO taO wOa xap dat mak tkUj t hY  
vOq t se sOm wEn dk m de dEm vek t hEm pkaj\ s.nqj\ vek m va mOj\  
puOj\ sAQ kUq\ tK m aUk pjz t hem mUn nkin tKaY m aUk nkUq s nqaY  
naO t kUq\ tkaY m hYt t lm a de m khOj\ tAm trj t daY m daY cko t pjz  
kah jEu m s jO hEm daH t kAh qUen m yk...!! T mOq m dOz va hjU va t vs  
m pjO la 2 nq xa la h.p m nKE kOn duOq m yk da hEm kOa t.t se hem kOn  
pkjEn m nUa tkUj tkj h.p m nKE t la tke jEu^ m rAt nkjU nkUq lOj\ taj m  
hEm pjz trAn trOq tkj dUq traz taO...!!!  
P/s kOa 1sU rAh\ nkE...!!!

Thích · Bình luận · Chia sẻ

## Báo mạng?

KINH DOANH TÀI CHÍNH DOANH NGHIỆP MUA SẮM ĐẦU TƯ

Giá vàng Hồi phục kinh tế Xuất khẩu gạo

## Báo Anh: Việt Nam vượt Trung Quốc, trở thành điểm đầu tư hấp dẫn ở châu Á

15/01/2021 22:33 GMT+7



14



0



TTO - Báo cáo công bố ngày 14-1 của Đơn vị phân tích kinh tế (EIU) thuộc tạp chí The Economist cho thấy Việt Nam đang vươn lên nhanh chóng, trở thành trung tâm sản xuất thay thế Trung Quốc trong chuỗi cung ứng giá rẻ.

- South China Morning Post: ASEAN có thể học hỏi Việt Nam cách thu hút FDI
- 'Thời điểm thuận lợi cho Việt Nam thu hút có chọn lọc dòng vốn FDI'
- Thành công xưởng mới của thế giới, VN cần làm gì tiếp theo?

# Khai thác dữ liệu

- Nguồn báo mạng: Tuổi trẻ Online (<https://tuoitre.vn/>)
  - Có cấu trúc rõ ràng > Dễ parse
  - Dữ liệu đa dạng, nhiều

## TIN MỚI NHẤT

Thứ 5, ngày 7 tháng 1 năm 2021

Thông tin tuyển sinh Chuyển đổi số Vaccine ngừa COVID-19

Hôm nay  
07/01



THẾ GIỚI

### Chuyến gì sắp diễn ra tại Quốc hội Mỹ?

TTO - Nếu có nghị sĩ từ cả Hạ viện và Thượng viện Mỹ nhất trí phân đối phiếu đại cử tri tại một bang nào đó trong phiên kiểm phiếu đại cử tri ngày 6-1, đây sẽ là lần thứ ba việc...

- Phó tổng thống Mỹ hoan nghênh các nghị sĩ phân đối kết quả bầu cử theo luật pháp

Hôm nay  
07/01



SỨC KHỎE

### Vương quốc Anh ghi nhận hơn 1.000 người chết một ngày vì COVID

TTO - Vương quốc Anh ngày 6-1 ghi nhận thêm 1.041 người chết vì virus corona, cũng là lần đầu tiên kể từ tháng 4-2020 số người chết trong 24 giờ vì COVID-19 ở đây vượt...

- Thêm nhiều nơi phát hiện biến thể corona mới, Anh cho tiêm vắc xin COVID-19 thứ 2

Hôm qua  
06/01



PHÁP LUẬT

### Gã xe ôm chở khách nữ ra cánh đồng vắng cướp tài sản, hiếp dâm

TTO - Trần Văn Tịnh nhận chở một khách nữ đi từ Hà Nội về Bắc Ninh, tuy nhiên lại cầm tay khách nữ đi lạc đường cướp tài sản, hiếp dâm.

- 1 ca COVID-19 mới, phát hiện sau 3 lần xét nghiệm, hơn 20 ngày nhập cảnh

Hôm qua  
06/01



THỂ THAO

### Phố núi chào đón 'người nhà' Kiatisak

TTO - Chiều muộn 6-1, sân bay Pleiku huyền ảo, nhộn nhịp hơn khi chuyến bay VN 1426 đưa HLV Kiatisak và các cầu thủ Hoàng Anh Gia Lai trở về từ giải bóng đá tiền V-League...

- HLV Kiatisak: 'Hàng thủ là điểm yếu của Hoàng Anh Gia Lai'

Xem thêm

Xem theo thời gian

## CƠ HỘI MUA SẮM



# Khai thác dữ liệu

Quy trình lấy dữ liệu:

1. Kéo xuống cuối trang
2. Bấm “Xem thêm” và quay lại bước 1
3. Khi làm đủ nhiều thì Parse HTML

▷ Sử dụng Selenium

▷ 480 bài/2 phút

# Khai thác dữ liệu

- Thời gian thu thập tạm ổn
- Trang web dạng Infinity scroll
  - ▷ Bài báo xuất hiện khi bấm xem thêm từ đâu ra?
  - ▷ Trình duyệt request lên, server trả về, trình duyệt render thêm thông tin mới



TTO - Thấy lực lượng chức năng, lợi dụng thời tiết sương mù, các công dân nhập cảnh trái phép bỏ chạy nhưng bị lực lượng bộ đội biên phòng phát hiện và bắt giữ.

- 1 ca COVID-19 mới, phát hiện sau 3 lần xét nghiệm, hơn 20 ngày nhập cảnh

THỂ THAO

### Phố núi chào đón 'người nhà' Kiatisak

TTO - Chiều muộn 6-1, sân bay Pleiku huyền ảo, nhộn nhịp hơn khi chuyến bay VN 1426 đưa HLV Kiatisak và các cầu thủ Hoàng Anh Gia Lai trở về từ giải bóng đá tiền V-League...

- HLV Kiatisak: 'Hàng thủ là điểm yếu của Hoàng Anh Gia Lai'

THỂ GIỚI

### Trung Quốc ngăn dòng Mekong từ 31-12, tới ngày 5-1-2021 mới thông báo?

TTO - Ủy hội sông Mekong (MRC) và chính quyền Thái Lan hôm nay 6-1 cho biết Trung Quốc đã thông báo về việc ngăn dòng chảy tại đập Cảnh Hồng ở thượng nguồn sông...

- Nông dân Lào, Thái mất mùa, Trung Quốc lại chặn đập Cảnh Hồng 'bảo trì lưới điện'

KINH DOANH

### Xuất khẩu lâm sản là cứu tinh của ngành nông nghiệp

TTO - Bộ trưởng Nguyễn Xuân Cường cho rằng kim ngạch xuất khẩu lâm sản đã vượt kế hoạch, đạt 13,17 tỉ USD, và là cứu tinh trong nhiệm vụ xuất khẩu của cả ngành nông...

- Xuất khẩu gỗ: hàng ngàn container hàng bị tồn tại các cảng biển

THỜI SỰ

### Quân khu 5 tặng bằng khen cho báo Tuổi Trẻ

TTO - Chiều 6-1, Bộ tư lệnh Quân khu 5 tổ chức buổi gặp mặt và tuyên dương, khen thưởng các cơ quan báo chí có nhiều đóng góp tuyên truyền quân sự, quốc phòng Quân khu...

- Chủ tịch Quốc hội thăm Bộ tư lệnh Quân khu 5

KINH DOANH

### Wall Street Journal: Bắc Kinh ép Jack Ma chia sẻ dữ liệu người dùng của Ant Group

TTO - Chính quyền Trung Quốc đang cố gắng buộc Jack Ma làm cái việc mà tỉ phú này vẫn khước từ lâu nay: chia sẻ dữ liệu tín dụng người dùng của Ant Group.

- Đăng sau sự kiện Jack Ma 'mất tích'

PHÁP LUẬT

ElementsConsoleSourcesNetworkPerformanceMemoryApplicationSecurity»

Search

Filter☐ Hide data URLsAllXHRJS CSS Img Media Font Doc WS Manifest Other

☐ Has blocked cookies☐ Blocked Requests

500 ms1000 ms1500 ms2000 ms2500 ms3000 ms

Name

☐ trang-2.htm

☐ getcount-comment.api

Hôm qua 06/01

[Thế giới](#)

[Trung Quốc ngăn dòng Mekong từ 31-12, tới ngày 5-1-2021 mới thông báo?](#)

TTO - Ủy hội sông Mekong (MRC) và chính quyền Thái Lan hôm nay 6-1 cho biết Trung Quốc đã thông báo về việc ngăn dòng chảy tại đập Cảnh Hồng ở thượng nguồn sông Mekong trong 20 ngày, từ 5 đến 24-1.

[Nông dân Lào, Thái mất mùa, Trung Quốc lại chặn đập Cảnh Hồng 'bảo trì lưới điện'](#)

Hôm qua 06/01

[Kinh doanh](#)

[Xuất khẩu lâm sản là cứu tinh của ngành nông nghiệp](#)

TTO - Bộ trưởng Nguyễn Xuân Cường cho rằng kim ngạch xuất khẩu lâm sản đã vượt kế hoạch, đạt 13,17 tỉ USD, và là cứu tinh trong nhiệm vụ xuất khẩu của cả

2 / 24 requests11.2 kB / 11.2 kB tr



# Khai thác dữ liệu

- Request đường link:

<https://tuoitre.vn/timeline/0/trang-{index}.htm>

- Trong file htm trả về ở trên chứa link tới 20 bài báo
  - ▷ 400 bài/phút và không cần phải mô phỏng trình duyệt!
  - ▷ Khoảng 800.000 bài sau 2 ngày

# Khám phá

- Làm song song với khâu thu thập
  - ▷ Thử nghiệm trên 52.000 bài trước rồi truyền thông lại kết quả cho khâu thu thập
- Dữ liệu có 5 thuộc tính:
  - links: đường link gốc của bài báo
  - title: tựa đề bài báo
  - description: mô tả ngắn về bài báo
  - content: nội dung chính
  - class: chuyên mục của bài báo trên web

# Khám phá

- 5 thuộc tính, tất cả đều là chuỗi ký tự

	links	title	description	content	class
0	<a href="https://tuoitre.vn/tong-thong-trump-xac-nhan-k...">https://tuoitre.vn/tong-thong-trump-xac-nhan-k...</a>	Tổng thống Trump xác nhận không dự lễ nhậm chức...	TTO - Tổng thống Mỹ Donald Trump đăng tweet ch...	Sau khi đưa ra cam kết sẽ đảm bảo chuyển giao ...	Thế giới
1	<a href="https://tuoitre.vn/dat-nuoc-dat-niem-tin-vao-n...">https://tuoitre.vn/dat-nuoc-dat-niem-tin-vao-n...</a>	Đất nước đặt niềm tin vào những học sinh xuất sắc	TTO - Tối 8-1, Thủ tướng Chính phủ Nguyễn Xuân...	Chia sẻ tại buổi lễ, Thủ tướng Nguyễn Xuân Phú...	Giáo dục
2	<a href="https://tuoitre.vn/luat-su-my-phan-bien-ong-tr...">https://tuoitre.vn/luat-su-my-phan-bien-ong-tr...</a>	Luật sư Mỹ phản biện: Ông Trump đâu có kêu ngư...	TTO - Trên chương trình Bill Hemmer Reports củ...	Ngày 6-1, tình trạng bạo lực đã xảy ra tại tòa...	Thế giới
3	<a href="https://tuoitre.vn/thanh-pho-phu-quoc-se-phat-...">https://tuoitre.vn/thanh-pho-phu-quoc-se-phat-...</a>	Thành phố Phú Quốc sẽ phát triển dựa trên 4 tr...	TTO - Tối 8-1 tại phường An Thới, chính quyền ...	Phát biểu tại buổi lễ công bố thành lập TP Phú...	Thời sự
4	<a href="https://tuoitre.vn/ong-trump-nguoi-ung-ho-toi-...">https://tuoitre.vn/ong-trump-nguoi-ung-ho-toi-...</a>	Ông Trump: Người ủng hộ tôi 'sẽ có tiếng nói t...	TTO - Trong nội dung đăng trên Twitter sau gần...	"75 triệu người Mỹ yêu nước vĩ đại đã bầu cho ...	Thế giới

# Khám phá

- Dữ liệu có trùng nhưng ít ▷ Do web update bài mới làm bài cũ bị đôn xuống trang sau gây trùng lúc thu thập

	links	title	description	content	class
48880	/dang-sai-lech-vu-dong-tam-facebook-chuong-m...	Đăng sai lệch vụ Đồng Tâm, Facebooker 'Chương ...	TTO - Ngày 20-1, Cơ quan cảnh sát điều tra Côn...	Facebooker "Chương May Mẩn" tên thật là Chung ...	Pháp luật
48879	/dang-sai-lech-vu-dong-tam-facebook-chuong-m...	Đăng sai lệch vụ Đồng Tâm, Facebooker 'Chương ...	TTO - Ngày 20-1, Cơ quan cảnh sát điều tra Côn...	Facebooker "Chương May Mẩn" tên thật là Chung ...	Pháp luật
307	https://tuoitre.vn/1-ca-covid-19-moi-phat-hien...	1 ca COVID-19 mới, phát hiện sau 3 lần xét ngh...	TTO - Sau nhiều ngày số mắc mới khá cao và đồn...	Theo Bộ Y tế, 1 ca mắc mới ngày 6-1 là ca nhập...	Sức khỏe
254	https://tuoitre.vn/1-ca-covid-19-moi-phat-hien...	1 ca COVID-19 mới, phát hiện sau 3 lần xét ngh...	TTO - Sau nhiều ngày số mắc mới khá cao và đồn...	Theo Bộ Y tế, 1 ca mắc mới ngày 6-1 là ca nhập...	Sức khỏe
179	https://tuoitre.vn/affordable-luxury-dong-san-...	'Affordable luxury' - dòng sản phẩm đột phá về...	Apec Mandala Wyndham Mũi Né đánh dấu bước đầu ...	Apec Group và thương hiệu bất động sản mang tí...	Cần biết
...	...	...	...	...	...
229	https://tuoitre.vn/vuong-quoc-anh-ghi-nhan-hon...	Vương quốc Anh ghi nhận hơn 1.000 người chết m...	TTO - Vương quốc Anh ngày 6-1 ghi nhận thêm 1....	Theo hãng tin Reuters, trong ngày 6-1, Vương q...	Sức khỏe
304	https://tuoitre.vn/wall-street-journal-bac-kin...	Wall Street Journal: Bắc Kinh ép Jack Ma chia ...	TTO - Chính quyền Trung Quốc đang cố gắng buộc ...	Báo Wall Street Journal nhận định Jack Ma có r...	Kinh doanh

# Khám phá

- Dữ liệu thiếu tập trung nhiều ở content
  - ▷ Lúc tiền xử lý xóa bất kỳ dòng chứa dữ liệu thiếu

---

```
links          0
title          244
description     292
content        908
class          0
dtype: int64
```

# Khám phá

Phân lớp toàn dữ liệu  
nhiều và phân bố rất  
bất cân xứng

	class	count
0	xã hội	107988
1	thể giới	78925
2	thể thao	69875
3	kinh doanh	53261
4	văn hóa	43199
...	...	...
123	kết nối	2
124	tâm nhìn	2
125	khám phá	2
126	chuyện thành phố	1
127	cuộc thi viết	1

128 rows x 2 columns

# Tiền xử lý

- Do dữ liệu lớn nên chia 2 giai đoạn:
  - Tiền xử lý thô: xóa dữ liệu trùng và thiếu
  - Tiền xử lý phức tạp: chuẩn hóa chữ, xóa ký tự đặc biệt,...

# Tiền xử lý

- Unicode có 2 kiểu: dựng sẵn và tổ hợp
- Ví dụ để hiển thị chữ ẽ gồm có 2 cách:
  - Kết hợp ê + dấu ˜ (tổ hợp là phương pháp cũ)
  - Dựng sẵn ẽ trong bản chữ unicode (dựng sẵn)

(source: <https://fontviet.com/khac-biet-giua-unicode-to-hop-va-unicode-dung-san/>)

▷ Chuẩn hóa unicode: sử dụng thư viện có sẵn của python



# Tiền xử lý

- Tiếng Việt có 2 kiểu gõ dấu
  - ▷ Chuẩn hóa dấu câu về kiểu cũ (dùng code có sẵn)

Cũ	Mới
òa, óa, ỏa, ỏa, ọa	oà, óá, oả, oã, ọạ
òe, óe, ỏe, ỏe, ọe	oè, oé, oẻ, oễ, ọẹ
ùy, úy, ủy, ỹy, ụy	uỳ, úý, uỷ, uỹ, ụy

(nguồn: Wikipedia)

# Tiền xử lý

- Tiếng Việt không thể tách từ đơn giản theo dấu cách được
  - Ví dụ: “Học sinh học sinh học”
    - Tách theo dấu cách: “học” x 3 + “sinh” x 2
    - Thực tế: “học sinh” + “học” + “sinh học”
- ▷ Sử dụng thư viện có sẵn (pyvi)

# Tiền xử lý

- Stopword (từ dừng) là các từ xuất hiện trong hầu hết văn bản nhưng không mang nặng ngữ nghĩa
- Ví dụ: dù anh cứ đi, em cũng kệ ▷ anh đi, em kệ
  - ▷ Sử dụng bộ stopwords có sẵn để lọc
  - ▷ Xóa ký tự đặc biệt (.,|@>...), chuyển về chữ in thường

# Tiền xử lý

- Số hóa dữ liệu: Chuyển văn bản (string) về vector TF-IDF (số)
- Chuyển phân lớp về số (dùng LabelEncoder) giúp tính mô hình học nhanh hơn
- Dùng phương pháp Chi bình phương để rút gọn đặc trưng của vector TF-IDF.

## TFIDF

For a term  $i$  in document  $j$ :

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

$tf_{i,j}$  = number of occurrences of  $i$  in  $j$

$df_i$  = number of documents containing  $i$

$N$  = total number of documents

# Tiền xử lý

- Data có 127 lớp ▷ Cần chọn lọc phân lớp
- Sau khi xem xét thì quyết định chọn 18 lớp

1. thời sự quốc tế	2. thời sự trong nước	3. du lịch
4. kinh doanh	5. giải trí	6. công nghệ
7. nhà đất	8. sức khỏe	9. giáo dục
10. khoa học	11. thể thao	12. văn hóa
13. pháp luật	14. yêu	15. xe
16. thời trang	17. nhịp sống trẻ	18. ăn gì

# Xây dựng mô hình và thử nghiệm

Các mô hình sẽ được nhóm thử nghiệm trên 5% dữ liệu:

- Naive Bayes
- Logistic Regression
- Logistic Regression dùng với kỹ thuật bagging
- Neuron Network
- Neuron Network dùng với kỹ thuật bagging

Do ít dữ liệu, có một số phân lớp dù được chọn nhưng buộc phải loại đi vì quá ít (ngưỡng  $\max(10, 0.1\% \text{ dữ liệu})$ )

# Naive Bayes (5% data)

Thời gian học: <1s

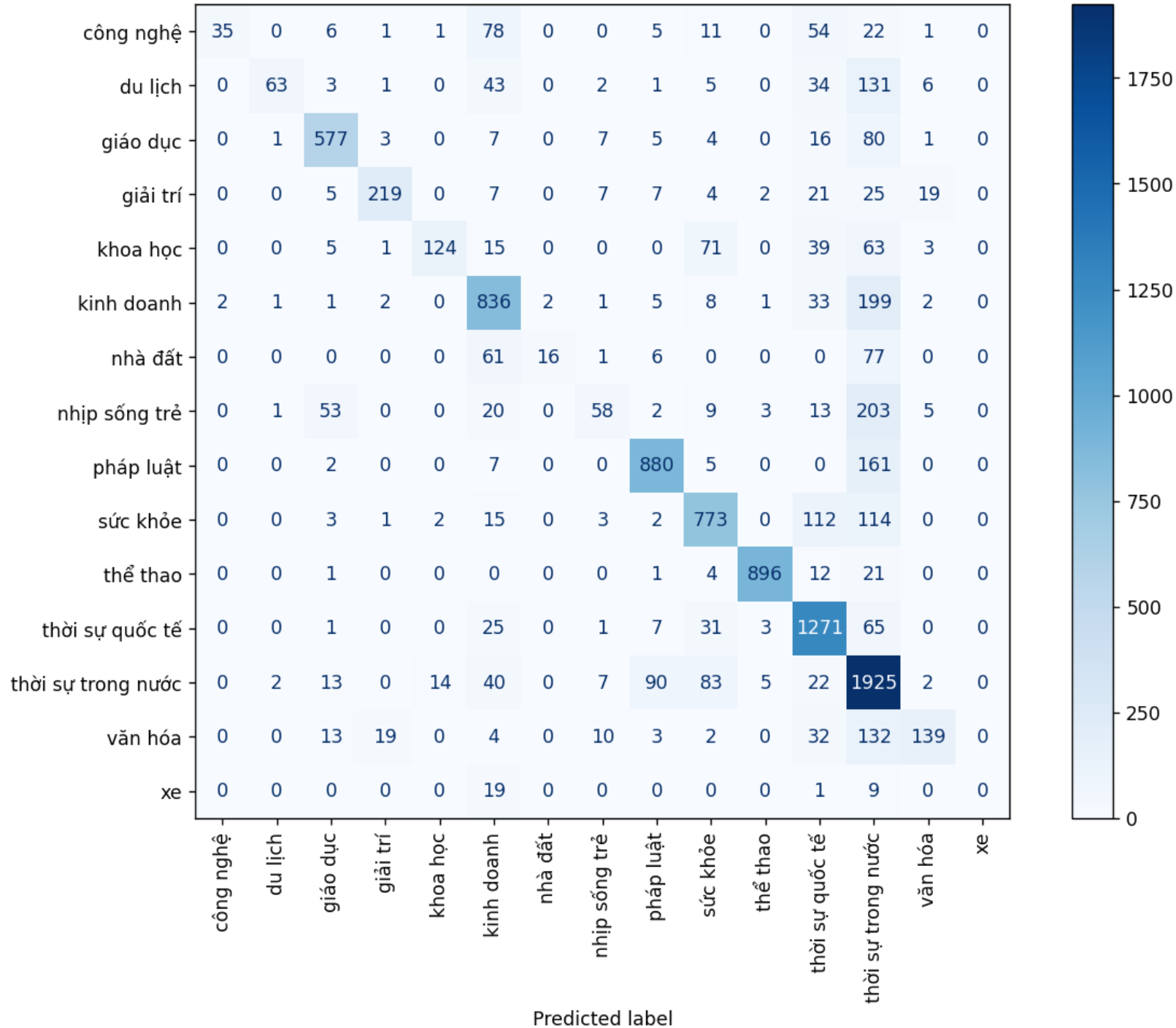
Đánh giá trên tập train:  
77.35%

Đánh giá trên tập validate:  
74.63%

▷ Predict có xu hướng tập  
trung vào class "thời sự  
trong nước"

▷ Class "xe" dự đoán  
sai 100%

True label



# Naive Bayes (5% data)

- Rất nhanh nhưng kết quả kém nhất
  - Predict có xu hướng tập trung vào class "thời sự trong nước"
  - Class "xe" dự đoán sai 100%
- ▷ Dễ hiểu vì:
- Tập dữ liệu khá nhỏ so với số lượng thuộc tính tạo bởi TF-IDF (hơn 10.000)
  - Phân lớp bị lệch khá là nặng về vài lớp chiếm đa số



# Logistic Regress\_ (5% data)

Thời gian học: 10-15s

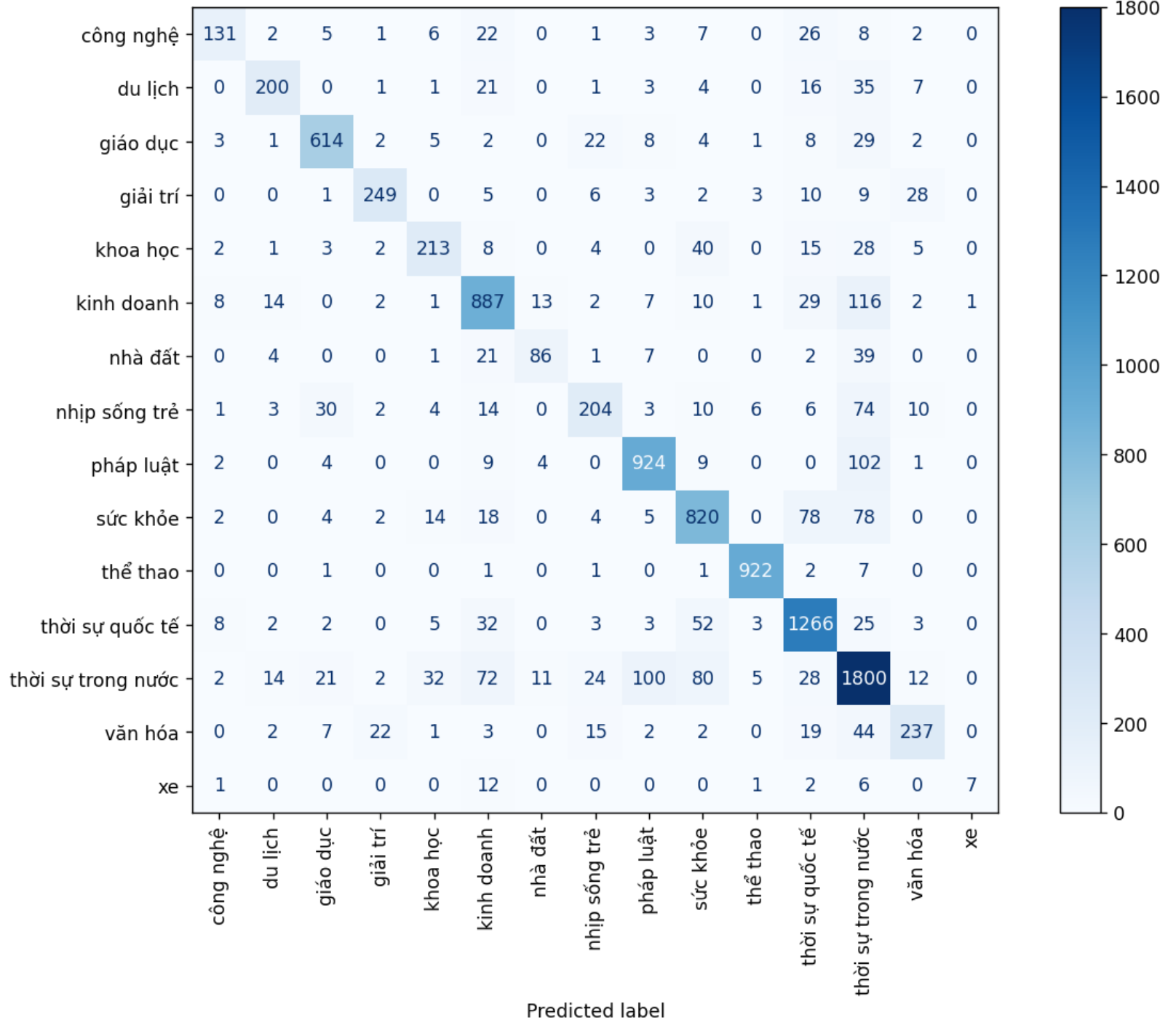
Đánh giá trên tập train:  
87.68%

Đánh giá trên tập validate:  
81.78%

▷ Chạy chậm hơn rất nhiều

▷ Dự đoán tốt hơn (không  
bất ngờ lắm)

True label



# Neuron Network (5% data)

Thời gian học: ~26s

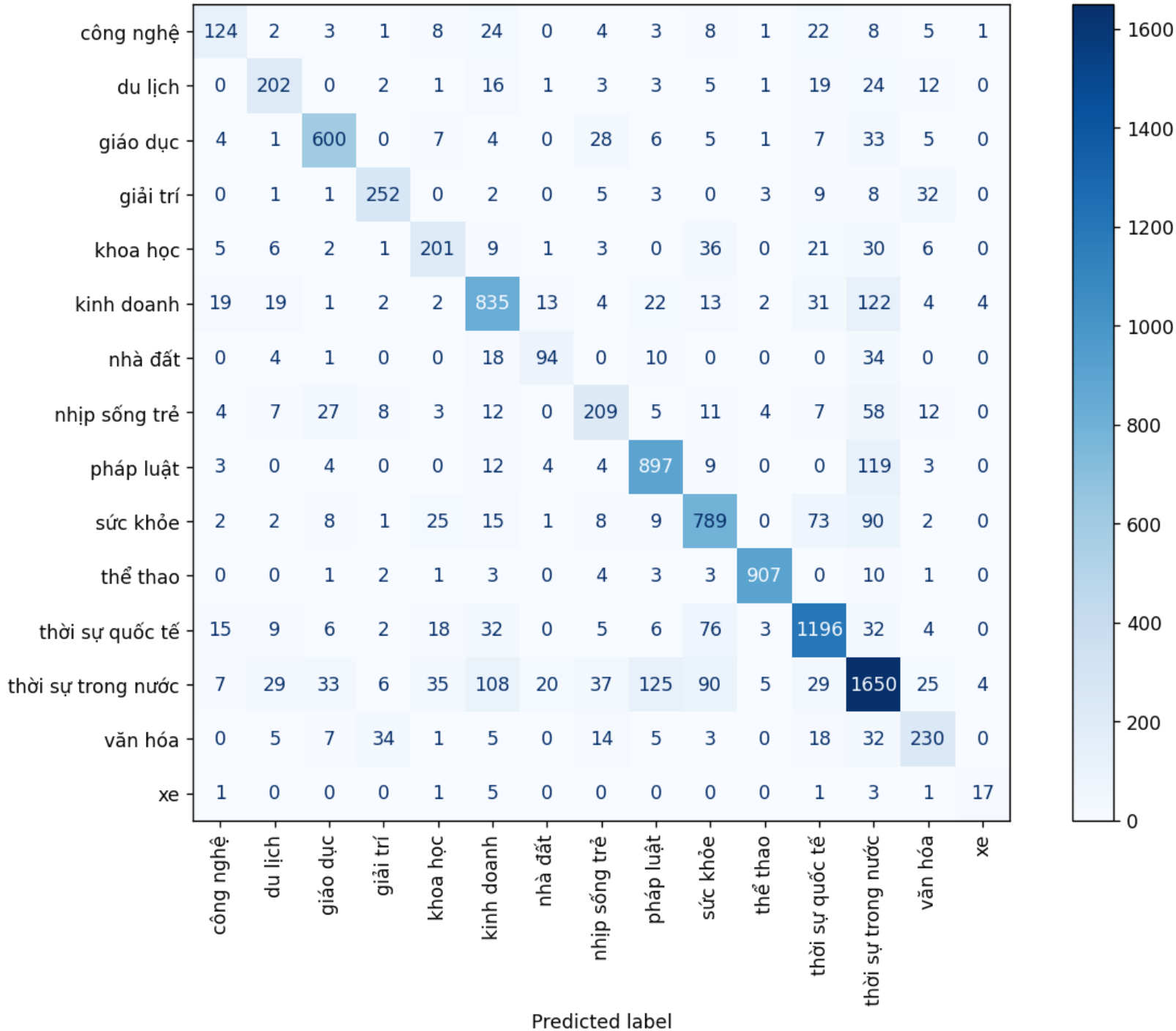
Đánh giá trên tập train:  
100%

Đánh giá trên tập validate:  
78.37%

▷ Chạy chậm (gấp 2 lần  
logistic) mà lại bị overfit

▷ Mô hình neuron network  
học "tốt quá" mà lại gặp dữ  
liệu khá thưa :v

True label



# Phương pháp Bagging

Từ một tập dữ liệu gốc, lấy  $n$  samples (kiểu tổ hợp lặp) và cho  $n$  mô hình học

▷ Tạo ra một model gộp mà khi dự đoán mẫu mới, kết quả dự đoán sẽ được lấy theo “phiếu phổ thông” của  $n$  mô hình

“The bagged classifier often has **significantly greater accuracy than a single classifier** derived from  $D$ , the original training data. It will not be considerably worse and is **more robust to the effects of noisy data and overfitting**. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers.”

– “*Data Mining Concepts and Techniques*” - J.Han (2012)

# Bagging Logistic Regress\_

## (5% data)

Thời gian học: ~1p30s

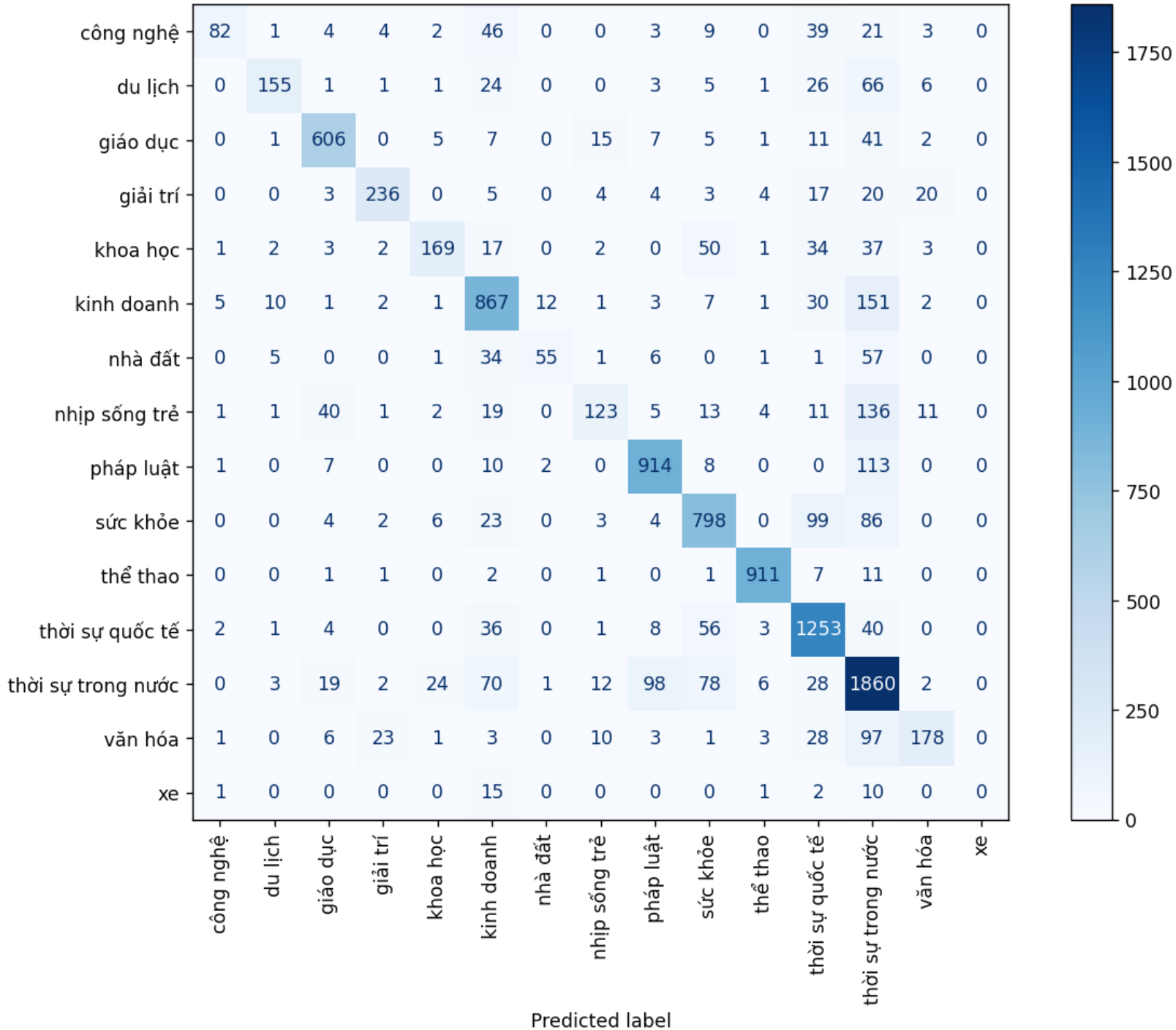
Đánh giá trên tập train:  
80.57%

Đánh giá trên tập validate:  
78.41%

▷ Thời gian chạy gấp lên  
xấp xỉ số lượng estimator  
(gấp 10)

▷ Tại lớp có số lượng ít như  
"xe" dự đoán sai hoàn toàn

True label



# Bagging Logistic Regression (5% data)

- Thời gian chạy gấp lên xấp xỉ số lượng estimator (gấp 10)
- Tại lớp có số lượng ít như "xe" dự đoán sai hoàn toàn
  - ▷ Bagging chia nhỏ dữ liệu ra cho 10 mô hình cùng học
  - ▷ Dữ liệu đã nhỏ sẵn nay còn nhỏ hơn. Việc bị fit vào một vài lớp chiếm đa số là không thể tránh khỏi

# Bagging Neuron Net\_ (5% data)

Thời gian học: ~40s

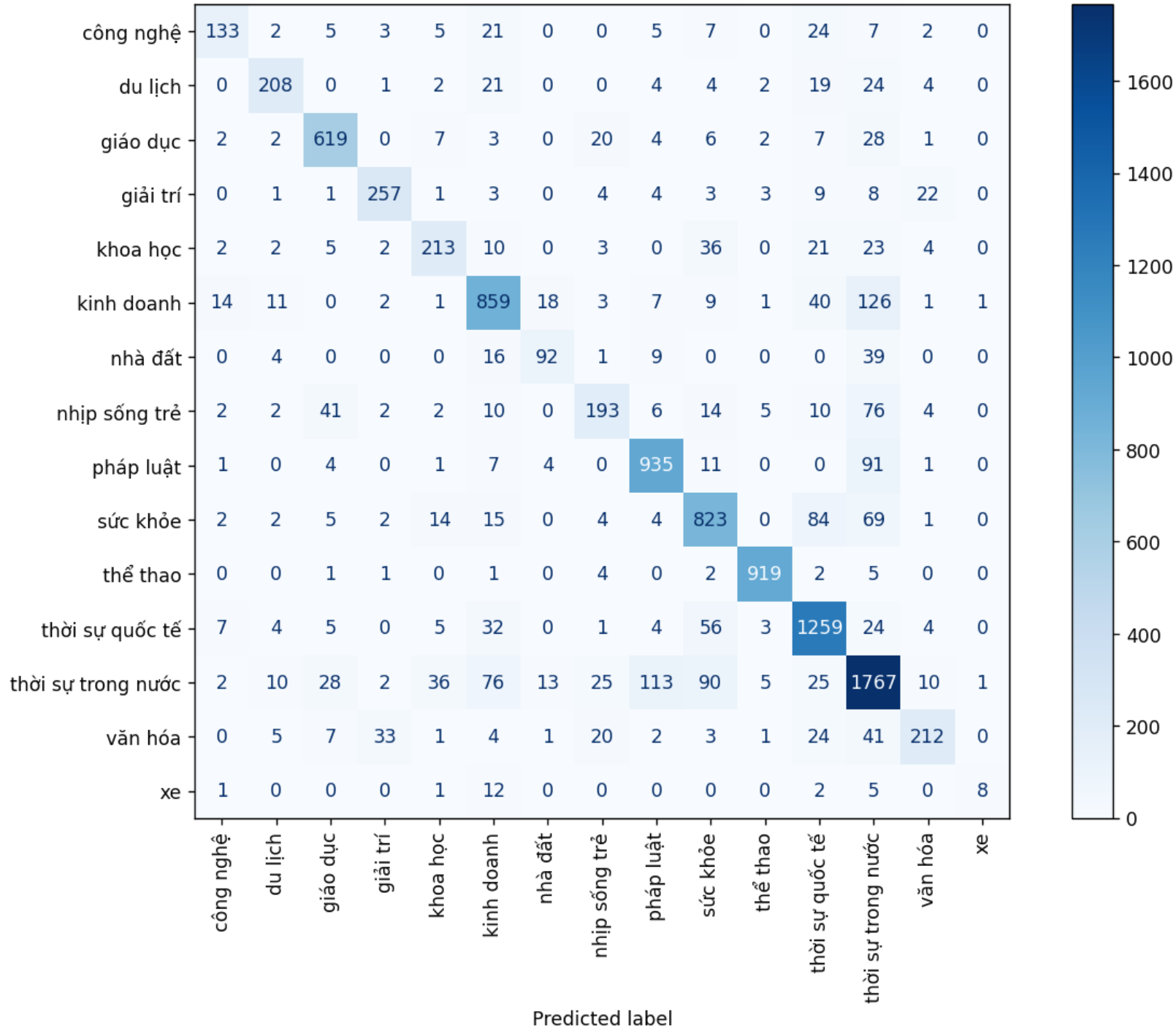
Đánh giá trên tập train:  
86.27%

Đánh giá trên tập validate:  
81.18%

▷ Tốc độ chạy chậm gấp  
đôi so với neuron network

▷ Tại lớp có số lượng ít như  
"xe" dự đoán chỉ kém  
neuron network thuần

True label



# Bagging Neuron Network (5% data)

- Tốc độ chạy chậm gấp đôi so với neuron network trước nhưng lại cho độ chính xác cao ngang mô hình logistic
- Tại lớp có số lượng ít như "xe" dự đoán chỉ kém neuron network thuần
  - ▷ Bagging biến bất lợi của neuron network trong trường hợp này là dễ bị overfit khi dữ liệu ít thành lợi thế
  - ▷ Lớp có số mẫu rất thấp như là "xe" được dự đoán tốt hơn, những lớp còn lại vẫn giữ được khả năng dự đoán chứ không bị giảm đi

# Xây dựng mô hình và thử nghiệm

- Do khối lượng dữ liệu lớn, không thể thử nghiệm được mô hình học SVM cũng như thử nghiệm nhiều siêu tham số khác nhau của mô hình Neuron Network
- Chọn 3 mô hình từ các thử nghiệm trước:
  - Naive Bayes
  - Logistic Regression
  - Bagging Neuron Network



# Thử nghiệm toàn dữ liệu (có chia tập tỉ lệ 70/30)

	Naive Bayes	Logistic Regression	Bagging NN
Thời gian chạy	2.6s	3p08s	12p34s
Đánh giá tập train	76.52%	81.18%	80.80%
Đánh giá tập validate	76.22%	79.73%	79.57%

- Kết quả gần tương tự như thử nghiệm trên 5%
- Bagging Neuron Network chạy nhanh hơn thử nghiệm cũ với phiên bản gốc (mất khoảng 1h train) nhưng để dự đoán 1 mẫu thì mất 30s (do sử dụng 30 model con)

# Demo model cuối cùng (train 100% data)

- Test 1: lời bài hát “Có như không có” của Hiền Hồ

Anh lại để lạc mất em rồi  
Lại để em ở một mình đành lòng anh sao?  
Lại để cô gái anh yêu phải khóc  
Em vẫn cam lòng và không than trách nửa lời  
...  
Ừ thì đã có, nhưng có như không mà thôi  
Nhưng có như không vậy thôi  
Anh chăm lo người ta mất rồi  
Thật lòng...

# Naive Bayes (Gud 🙌)

Naive Bayes Pred: ['yêu'] | Prob: 0.42

Prediction probabilities

yêu	0.42
nhịp sống trẻ	0.19
thời sự trong...	0.19
văn hóa	0.13
Other	0.06

NOT du lịch

du lịch

chăm\_lo 0.01  
đau 0.01  
mặt\_trời 0.01  
biển 0.01  
gương 0.01  
trách 0.01  
khóc 0.01  
hoa 0.00  
sống 0.00  
vui 0.00

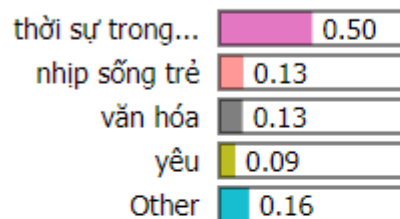
## Text with highlighted words

lạc một\_mình đành\_lòng gái yêu khóc cam than  
trách nửa hoa ánh mặt\_trời vắng đời vui biển sóng  
gương nát tan gương vỡ lành chăm\_lo người\_ta đau  
hoa ánh mặt\_trời vắng đời vui biển sóng dạt\_dào nát  
tan đau chăm\_lo người\_ta

# Logistic Regression (😄)

LogReg Pred: ['thời sự trong nước'] | Prob: 0.5

Prediction probabilities



NOT du lịch

du lịch

hoa 0.01  
người\_ta 0.01  
ánh 0.01  
biển 0.01  
đời 0.00  
vui 0.00  
gương 0.00  
đau 0.00  
chăm\_lo 0.00  
khóc 0.00

## Text with highlighted words

lạc một\_mình đành\_lòng gái yêu khóc cam than  
trách nửa hoa ánh mặt\_trời vắng đời vui biển sóng  
gương nát tan gương vỡ lành chăm\_lo người\_ta đau  
hoa ánh mặt\_trời vắng đời vui biển sóng dạt\_dào nát  
tan đau chăm\_lo người\_ta

# Bagging NN (🐼x30 > 🐼)

NN Pred: ['thời sự trong nước'] | Prob: 0.45

Prediction probabilities

thời sự trong...	0.45
nhịp sống trẻ	0.14
yêu	0.12
giáo dục	0.07
Other	0.21

NOT du lịch

đời 0.01  
chăm\_lo 0.01  
biển 0.01  
hoa 0.01  
gương 0.01  
ánh 0.00  
tan 0.00  
mặt\_trời 0.00  
vắng 0.00  
cam 0.00

du lịch

## Text with highlighted words

lạc một\_mình đành\_lòng gái yêu khóc cam than  
trách nửa hoa ánh mặt\_trời vắng đời vui biển sóng  
gương nát tan gương vỡ lành chăm\_lo người\_ta đau  
hoa ánh mặt\_trời vắng đời vui biển sóng dạt\_dào nát  
tan đau chăm\_lo người\_ta

# Demo model cuối cùng (train 100% data)

- Test 2: bài post bình luận chính trị quốc tế (lấy từ group “Maybe You Missed This F\*\*\*king News” trên Facebook)

Bài phân tích của The Atlantic: Giới lãnh đạo Trung Quốc có khi lại muốn Trump tái đắc cử hơn chúng ta vẫn nghĩ

Tác giả bài viết: Michael Schuman, đăng ngày 7 tháng 7 năm 2020

Link bài gốc:

<https://www.theatlantic.com/.../why-china-wants.../613864/>

Cũng như tất cả mọi người dân trên lãnh thổ quốc gia và trên thế giới, các nhà lãnh đạo của Trung Quốc hẳn cũng đang dõi theo kết quả của chiến dịch tranh cử khốc liệt đang diễn ra tại Hợp chủng quốc Hoa Kỳ và đang bồn chồn bồn khoăn không biết kết quả sẽ có tác động như thế nào đến họ. Sau bốn năm trời kèn cựa với Donald Trump, dân Trung chắc nên đếm từng tháng, từng tuần, từng ngày hay từng phút cho đến cuộc bầu cử vào tháng 11, hy vọng rằng ứng viên Dân chủ mềm dẻo hơn sẽ tiếp quản Nhà Trắng, đúng chứ? Đó chắc chắn là điều mà Trump tin. Người Trung Quốc, ông tweet, “đang tuyệt vọng mong Sleepy Joe Biden đắc cử để họ có thể tiếp tục lừa dối nước Mỹ, như họ đã làm hàng thập kỉ trước, cho tới khi tôi xuất hiện!”

Điều này chưa chắc đã đúng. Theo quan điểm của Bắc Kinh thì, trong khi một tổng thống đảng Dân chủ có thể tái thiết lập chính sách ngoại giao dễ đoán hơn của Hoa Kỳ, nó vẫn có thể không có lợi cho Trung Quốc. Trên thực tế, Trump tái đắc cử thêm 4 năm nữa – dù có vẻ sẽ đi kèm với phiền toái và tranh cãi – có khi lại là một cơ hội béo bở cho Trung Quốc mở rộng phạm vi ảnh hưởng tại Đông Á và trên thế giới.

# Naive Bayes (Gud Gud 🙌🙌)

Naive Bayes Pred: ['thời sự quốc tế'] | Prob: 1.0

Prediction probabilities

thời sự quốc tế	1.00
công nghệ	0.00
kinh doanh	0.00
văn hóa	0.00
Other	0.00

NOT du lịch

trump  
0.00  
dân\_chủ  
0.00  
đảng  
0.00  
tổng\_thống  
0.00  
bầu\_cử  
0.00  
mỹ  
0.00  
đàm\_phán  
0.00  
kì  
0.00  
ta  
0.00  
trung  
0.00

du lịch

## Text with highlighted words

phân\_tích the atlantic giới lãnh\_đạo trung quốc  
trump tái đắc\_cử dân lãnh\_thở quốc\_gia thế\_giới  
lãnh\_đạo trung quốc hẳn dôi kết\_quả chiến\_dịch  
tranh\_cử khốc\_liệt diễn hợp chủng quốc hoa kỳ  
bồn\_chồn bản\_khoản kết\_quả tác\_động bốn trời  
kèn\_cựa donald trump dân trung đếm tuần phút  
bầu\_cử 11 hy\_vọng ứng\_viên dân\_chủ mềm\_dẻo  
tiếp\_quản trắng trump trung quốc tweet tuyệt\_vọng  
mong sleepy joe Biden đắc\_cử lừa\_dối mỹ làm\_hàng  
thập\_kỉ quan\_điểm bắc kinh tổng\_thống đảng  
dân\_chủ tái\_thiết\_lập chính\_sách ngoại\_giao đoán  
hoa kì lời trung quốc trump tái đắc\_cử 4 về đi kèm

# Logistic Regression (👁️📝)

LogReg Pred: ['thời sự quốc tế'] | Prob: 1.0

Prediction probabilities

thời sự quốc tế	<div><div></div></div>	1.00
kinh doanh	<div><div></div></div>	0.00
thời sự trong...	<div><div></div></div>	0.00
công nghệ	<div><div></div></div>	0.00
Other	<div><div></div></div>	0.00

NOT du lịch

tổng\_thống  
0.00  
trung\_quốc  
0.00  
đảng  
0.00  
chính\_quyền  
0.00  
kinh  
0.00  
tuyên\_bố  
0.00  
lãnh\_đạo  
0.00  
mỹ  
0.00  
trung  
0.00  
thế\_giới  
0.00

du lịch

## Text with highlighted words

phân\_tích the atlantic giới lãnh\_đạo trung quốc trump tái đắc\_cử dân lãnh\_thở quốc\_gia thế\_giới lãnh\_đạo trung quốc hần dỗi kết\_quả chiến\_dịch tranh\_cử khốc\_liệt diễn hợp chủng quốc hoa kỳ bồn\_chồn bản\_khoản kết\_quả tác\_động bốn trời kền\_cựa donald trump dân trung đếm tuần phút bầu\_cử 11 hy\_vọng ứng\_viên dân\_chủ mềm\_dẻo tiếp\_quản trắng trump trung quốc tweet tuyệt\_vọng mong sleepy joe biden đắc\_cử lừa\_dối mỹ làm\_hàng thập\_kỷ quan\_điểm bắc kinh tổng\_thống đảng dân\_chủ tái\_thiết\_lập chính\_sách ngoại\_giao đoán hoa kỳ lợi trung quốc trump tái đắc\_cử 4 về đi kèm



# Bagging NN (🙏🙏🙏)

NN Pred: ['thời sự quốc tế'] | Prob: 1.0

Prediction probabilities

thời sự quốc tế	1.00
thời sự trong...	0.00
giải trí	0.00
công nghệ	0.00
Other	0.00

NOT du lịch

tổng\_thống 0.00  
trump 0.00  
đắc\_củ 0.00  
ta 0.00  
kiên\_nhẫn 0.00  
dẫn\_đầu 0.00  
bắc 0.00  
êm\_ả 0.00  
dòng 0.00  
quốc\_tế 0.00

du lịch

trump tại đắc\_củ dân nam\_thứ quốc\_gia thế\_giới  
lãnh\_đạo trung quốc hẳn đổi kết\_quả chiến\_dịch  
tranh\_cử khốc\_liệt diễn hợp chủng quốc hoa kỳ  
bồn\_chồn bản\_khoản kết\_quả tác\_động bốn trời  
kèn\_cựa donald trump dân trung đếm tuần phút  
bầu\_cử 11 hy\_vọng ứng\_viên dân\_chủ mềm\_dẻo  
tiếp\_quản trắng trump trung quốc tweet tuyệt\_vọng  
mong sleepy joe Biden đắc\_củ lừa\_dối mĩ làm\_hàng  
thập\_kỉ quan\_điểm bắc kinh tổng\_thống đảng  
dân\_chủ tái\_thiết\_lập chính\_sách ngoại\_giao đoán  
hoa kì lợi trung quốc trump tái đắc\_củ 4 về đi kèm  
phiền\_toái tranh\_cãi béo\_bở trung quốc mở\_rộng  
phạm\_vi ảnh\_hưởng đông\_thế\_giới tất\_nhiên  
nhân\_vật trung quốc kết\_quả liệu quan\_điểm

# Demo model cuối cùng (train 100% data)

- Test 3: bài post anti  
(lấy từ group “Nói Không Với  
Hoa Hậu Đạo Lý” trên  
Facebook)

Series part 1:

ĐẠI SỨ NHAM HIỂM

Ban đầu tính đặt tên Đại Sứ Thảo Mai, nhưng cái tên này không thể nào xứng đáng với cô HHCG này được. Hai từ nham hiểm nó thể hiện rõ trong bản chất con người này. Có lẽ, các bạn đã xem series trước về sự mất dạy của Y đối với nghệ sĩ đàn anh của mình. Thì ngay sau đó, Y tham gia HHCG mà không qua bất kỳ một cuộc thi tuyển chọn nào.

Lòng dạ con người này muốn thoát xác từ Singer Gen Ni Phơ sang Hoa Hậu “thiếu chữ” Chuyển Giới. Một bước lên mây bằng cái danh hoa hậu “quốc tế”, Y và ekip mua tất cả các bài báo và tung hô như vừa được Miss Universe. Chiêu bài này quá quen thuộc trong scandal mới gần đây, đó là tẩy trắng sạch sẽ scandal mất dạy trước đó.

Hơi lạc đề một tí, đáng lẽ cuộc thi đó như các bạn biết vương miện đáng lẽ thuộc về anh phiên dịch mặc dù câu trả lời của Y và phiên dịch đều trật lất câu hỏi ban đầu của giám khảo đưa ra ☐ Và nên nhớ, vì Y không được bất kỳ tổ chức nào đưa đi thi, nên Y tham gia với tư cách là cá nhân chứ không phải đại diện Việt Nam tham gia chương trình quốc tế nhé.

Có cuộc thi quốc tế nào mà tự dưng bạn ở Thái Lan đăng ký vài ngày xong bảo Y vượt qua nhiều thí sinh khác để được chọn????? Đúng cái chương trình tả phỉ lù đến sợ.

Quay trở lại vấn đề, sau khi đoạt giải năm 2018, Y trở thành giám đốc chương trình để tìm kiếm tài năng cho chương trình này luôn =)). Vậy cũng hiểu động cơ của chương trình và Y đã thoả thuận như thế nào để được giải và đem cái chương trình này về rồi hen. Với tiêu chí, một vương quốc chỉ có một vua, Y dùng tất cả các chiêu bài để có thể dìm

# Naive Bayes (Hmm 😊)

Naive Bayes Pred: ['văn hóa'] | Prob: 0.75

Prediction probabilities

văn hóa	0.74
giải trí	0.23
nhịp sống trẻ	0.02
yêu	0.00
Other	0.00

NOT du lịch

thí_sinh	0.00
hoa_hậu	0.00
tập	0.00
thi	0.00
miss	0.00
vương_miện	0.00
universe	0.00
tư_cách	0.00
lù	0.00
quốc_tế	0.00

du lịch

## Text with highlighted words

series part 1 đại\_sứ nham\_hiểm ban\_đầu đại\_sứ thảo mai không\_thể\_nào xứng\_đáng hcg hai nham\_hiểm thể\_hiện bản\_chất con\_người có\_lẽ series mất\_dạy y nghệ\_sĩ đàn\_anh y tham\_gia hcg thi\_tuyển\_chọn lòng\_dạ\_con\_người thoát xác singer gen ni phơ hoa\_hậu chữ giới mây danh hoa\_hậu quốc\_tế y ekip mua báo tung\_hồ miss universe chiêu\_bài quen\_thuộc scandal tẩy trắng sạch\_sẽ scandal mất\_dạy hơi lạc\_đề một\_tí thi vương\_miện phiên\_dịch mặc\_dù câu trả\_lời y phiên\_dịch trật\_lất câu ban\_đầu giám\_khảo y tổ\_chức đi thi y tham\_gia tư\_cách đại diện viết nam tham gia chương trình

# Logistic Regression (🙄)

LogReg Pred: ['văn hóa'] | Prob: 0.59

Prediction probabilities

văn hóa	0.57
giải trí	0.37
nhịp sống trẻ	0.01
giáo dục	0.01
Other	0.04

NOT du lịch

hoa\_hậu 0.01  
chương\_trình 0.01  
thí\_sinh 0.00  
quốc\_tế 0.00  
tập 0.00  
nghệ\_sĩ 0.00  
đi 0.00  
miss 0.00  
thi 0.00  
thái 0.00

du lịch

## Text with highlighted words

series part 1 đại\_sứ nham\_hiểm ban\_đầu đại\_sứ thảo  
mai không\_thể\_nào xứng\_đáng hhcg hai nham\_hiểm  
thể\_hiện bản\_chất con\_người có\_lẽ series mất\_dạy y  
nghệ\_sĩ đàn\_anh y tham\_gia hhcg thi\_tuyển\_chọn  
lòng\_dạ\_con\_người thoát xác singer gen ni phơ  
hoa\_hậu chữ giới mây danh hoa\_hậu quốc\_tế y ekip  
mua báo tung\_hồ miss universe chiêu\_bài  
quen\_thuộc scandal tẩy trắng sạch\_sẽ scandal  
mất\_dạy hơi lạc\_đề một\_tí thi vương\_miện  
phiên\_dịch mặc\_dù câu trả\_lời y phiên\_dịch trật\_lất  
câu ban\_đầu giám\_khảo y tổ\_chức đi thi y tham\_gia  
tư cách đại diện việt nam tham gia chương\_trình

# Bagging NN (Yeah 👍)

NN Pred: ['văn hóa'] | Prob: 0.49

Prediction probabilities

giải trí	0.49
văn hóa	0.48
yêu	0.03
công nghệ	0.00
Other	0.00

NOT du lịch

tập  
0.00  
hoa\_hậu  
0.00  
chương\_trình  
0.00  
đi  
0.00  
thi  
0.00  
thí\_sinh  
0.00  
tham\_gia  
0.00  
quốc\_tế  
0.00  
thái  
0.00  
clip  
0.00

du lịch

## Text with highlighted words

series part 1 đại\_sứ nham\_hiểm ban\_đầu đại\_sứ thảo  
mai không\_thể\_nào xứng\_đáng hhcg hai nham\_hiểm  
thể\_hiện bản\_chất con\_người có\_lẽ series mất\_dạy y  
nghệ\_sĩ đàn\_anh y tham\_gia hhcg thi\_tuyển\_chọn  
lòng\_dạ\_con\_người thoát xác singer gen ni phơ  
hoa\_hậu chữ giới mây danh hoa\_hậu quốc\_tế y ekip  
mua báo tung\_hồ miss universe chiêu\_bài  
quen\_thuộc scandal tẩy trắng sạch\_sẽ scandal  
mất\_dạy hơi lạc\_đề một\_tí thi vương\_miện  
phiên\_dịch mặc\_dù câu trả\_lời y phiên\_dịch trật\_lất  
câu ban\_đầu giám\_khảo y tổ\_chức đi thi y tham\_gia  
tư cách đại diện việt nam tham\_gia chương\_trình

Tổng kết

# Đánh giá mô hình

- Độ chính xác không cao nhưng cũng đủ tốt (khoảng 75-80%)
- Sử dụng phương pháp bagging để tăng tốc độ chạy mô hình neuron network mà lại giảm overfit, giúp mô hình này đủ sức chạy trên toàn dữ liệu trong thời gian cho phép (< 20 phút)
- Không đủ khả năng thử nghiệm toàn bộ dữ liệu do thời gian xử lý khá lâu

# Đánh giá đồ án (chiêm nghiệm)


Chưa tốt ☹️

- Thời gian làm khá gấp rút do suy nghĩ đề tài muộn
- Không hoàn thành sớm được đồ án
- Không tìm được nguồn dữ liệu chuyên cho đề tài của nhóm
- Chưa có pipeline tổng quát



# Đánh giá đồ án (chiêm nghiệm)

Tốt 

- Thu thập được lượng lớn dữ liệu.
- Có khả năng tìm hiểu được các kiến thức cần thiết.
- Phân chia ra các quy trình khoa học dữ liệu riêng biệt và phân công mỗi thành viên nắm một quy trình riêng chứ không cùng làm chung (phần này nhóm không biết là điều xấu hay điều tốt, nhưng nhóm nghĩ khả năng cao là tốt ).

# Hướng phát triển

- Hoàn thiện pipeline tổng quát
- Tìm cách thử nghiệm nhiều hơn với mô hình
- Tìm hiểu các kỹ thuật, mô hình mới tốt hơn (BERT, Word2Vec,...)



 Kết thúc 

# Tham khảo

- Các notebook demo và bài tập của thầy (đặc biệt là bài tập 3 ♥)
- <https://www.scraping-bot.io/how-to-scrape-infinite-scroll-pages/>
- <https://prodevsblog.com/questions/128808/python-requests-requests-exceptions-toomanyredirects-exceeded-30-redirects/>
- <https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794>
- <https://nguyenvanhieu.vn/phan-loai-van-ban-tieng-viet/>
- <https://quan.hoabinh.vn/blog/2020/7/85-chuyen-doi-unicode-dung-san-to-hop-voi-python>
- <https://kipalog.com/posts/Gioi-thieu-tien-xu-ly-trong-xu-ly-ngon-ngu-tu-nhien>
- Document của scikit-learn, pandas, tqdm, pandarallel
- Và hằng hà sa số câu trả lời cho những câu hỏi ngu ngốc của nhóm tụi em trên stackoverflow và google