Text Classification

Nhóm 5

Thành viên

Họ và tên	Mã sinh viên	Lớp
Nguyễn Nhật Tân	22022530	K67-AI2
Nguyễn Tiến Dũng	22022644	K67-AI1
Ngô Việt Anh	22022508	K67-AI2

Bài toán

Bài toán: Classification - Phân loại tag của News

Data type: News

Input: Title

Output: Genre

Tiêu đề bài báo	Thể loại
Thủ tướng quy định trường hợp EVN được điều chỉnh giá bán lẻ điện bình quân	Kinh tế
9 thí sinh bị đình chỉ ngay đợt thi Đánh giá năng lực đầu tiên của ĐHQGHN	Giáo dục
Ô tô gầm cao cỡ nhỏ, dưới 600 triệu: Hyundai Venue giảm giá, đấu Kia Sonet	Xe
Bác sĩ viết đơn thuốc cẩu thả khiến bệnh nhân ngộ độc do uống quá liều	Sức khỏe
Bộ xử lý Apple M1, M2 và M3 dính lỗ hổng bảo mật không thể khắc phục	Công nghệ - game

Outline

01

Crawl & xử lý data

02

Data Annotation & Visualization

03

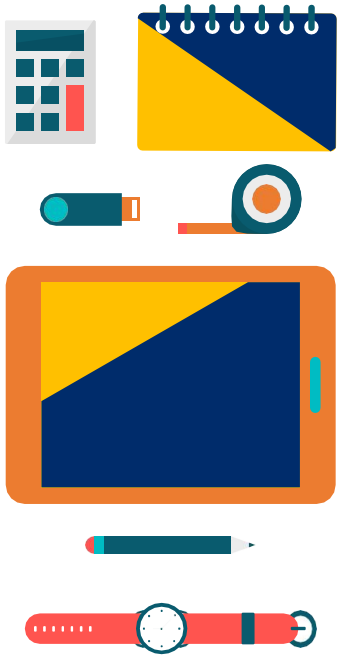
Models

04

Evaluation

05

Demo & Conclusion

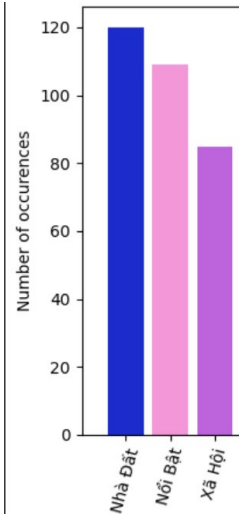


01

Crawl & xử lý data

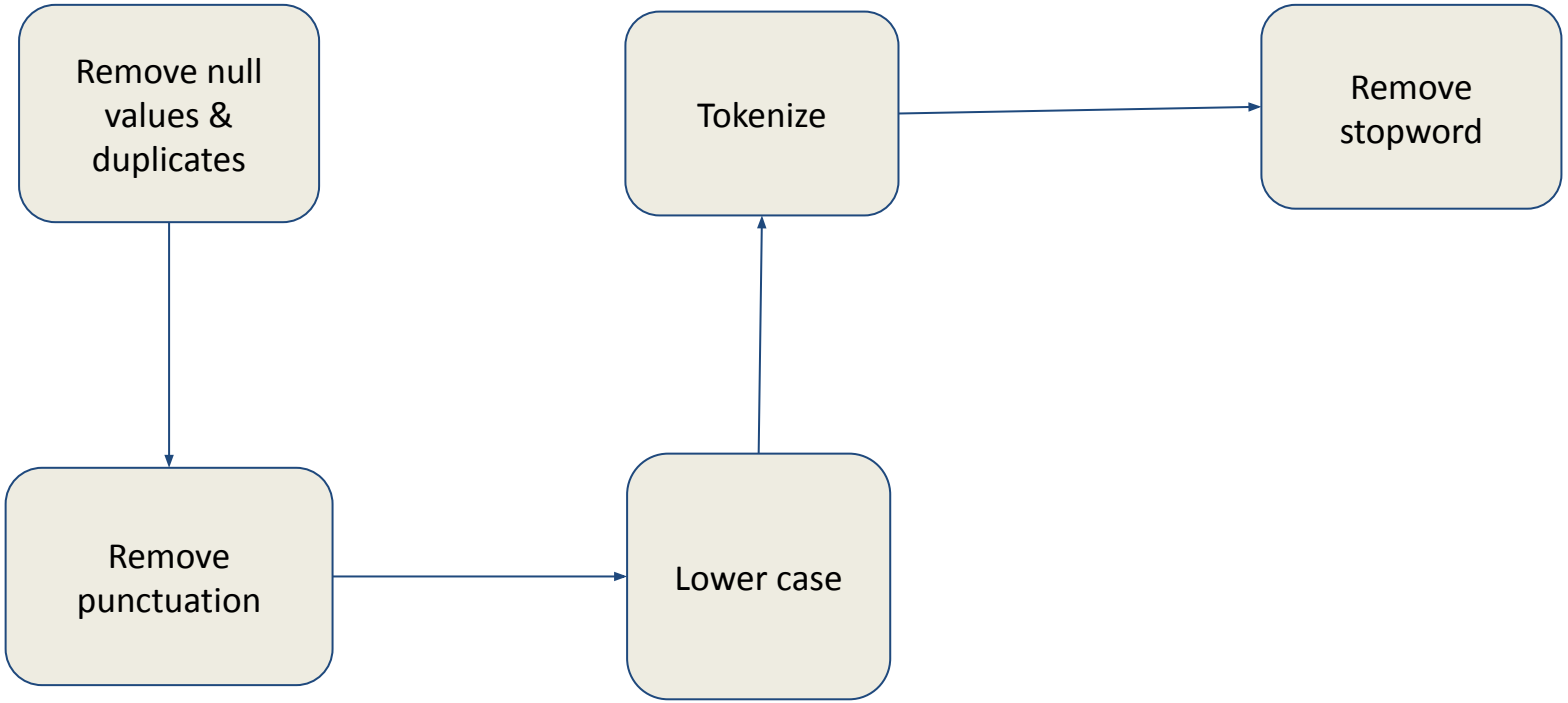
01 Crawl & data preprocessing

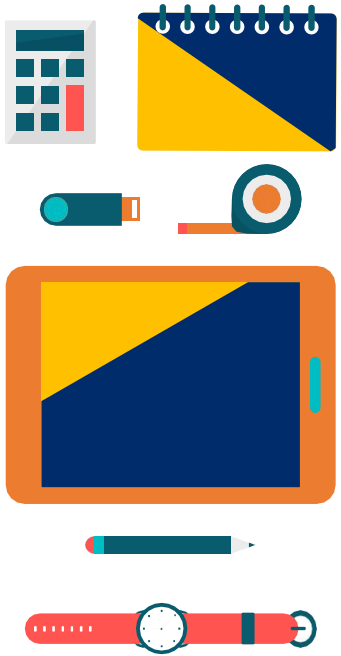
- Thử ở nhiều trang báo để chốt trang báo sẽ crawl
  => Data source: Báo Thanh niên
- Công cụ: Selenium + BeautifulSoup




```
8. Những căn bệnh dễ mắc phải khi trời nồm ẩm và cách phòng tránh
9. Phương tiện tăng nhanh so dự báo thiết kế, tuyến cao tốc Cam Lộ - La Sơn cần r
10. Vụ hiệu trưởng bị 'tố' chuyên quyền: Sở giáo dục chỉ đạo làm rõ
11. Nga tuyên bố sẽ đáp trả các cuộc tấn công xuyên biên giới
12. Nhận định Osasuna vs Real Madrid, 22h15 ngày 16/3: Khó có bất ngờ
13.
14. Chủ đề Năm Du lịch Quốc gia 2024 xây dựng trên âm hưởng hào hùng chiến thắng
15.
16. Thông tin bất ngờ vụ ô tô 'biến mất' ở thành phố Vinh được tìm thấy ở trạm y
17.
18. Thủ tướng Malaysia nói không nên quá kỳ vọng vào cuộc tìm kiếm xác MH370
19.
20. Điều động, bổ nhiệm nhiều vị trí cán bộ tại công an Hà Nội và 2 tỉnh, thành
```

01 Crawl & data preprocessing



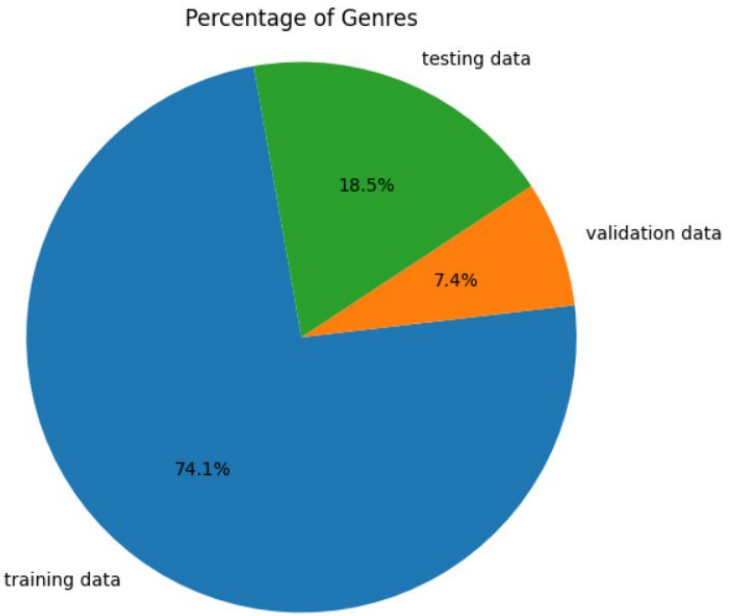


02

Data Annotation & Visualization

02 Data Annotation & Visualization

Data	Number of samples
Training	1000
Validation	100
Testing	250



Strategy:

- Gán nhãn thủ công.
- Sample ra 200 tập data, mỗi annotator sẽ gán nhãn cho tập đó rồi so sánh kết quả với nhau.
- Nếu có sự khác nhau => thảo luận để thống nhất kết quả.

=> Đạt được độ chính xác cao (đồng thuận 98%)

Challenge: Tốn nhiều tài nguyên nếu phải làm thủ công với bộ data lớn.

id	Title	Dũng	Tân	Việt Anh	tag thống nhất
625	Xe Trung Quốc 'nhái' Land Rover một thời tại Việt Nam rớt giá thê thảm	xe	xe	kinh tế	xe
717	Mitsubishi Xpander HEV 2024 thêm động cơ điện, trang bị 'xịn', giá hơn 600 triệu đồng	xe	xe	kinh tế	xe
1165	Các yếu tố giúp Samsung duy trì vị trí dẫn đầu trên thị trường bộ nhớ flash	công nghệ - game	cong-nghe-game	kinh tế	cong-nghe-game
1228	Doanh số iPhone giảm mạnh 24% tại Trung Quốc	kinh tế	cong-nghe-game	kinh tế	cong-nghe-game

Strategy:

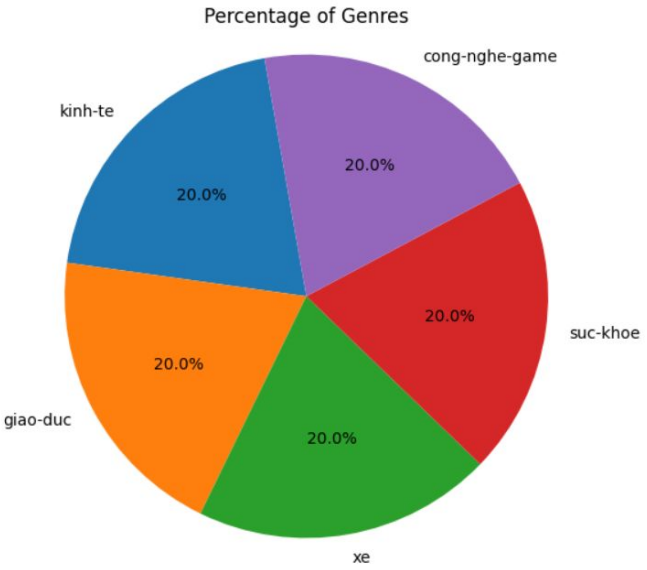
- Ý tưởng khác: Có 1 số từ đặc trưng, đại diện cho 1 số genre mà chỉ cần xuất hiện là có thể biết được genre. (eg: học sinh - ‘giáo dục’, ô tô - ‘xe’, ung thư - ‘sức khỏe’,...)
- Thống kê danh sách các từ, sau đó gán nhãn tự động cho các data có từ đó với nhãn tương ứng.

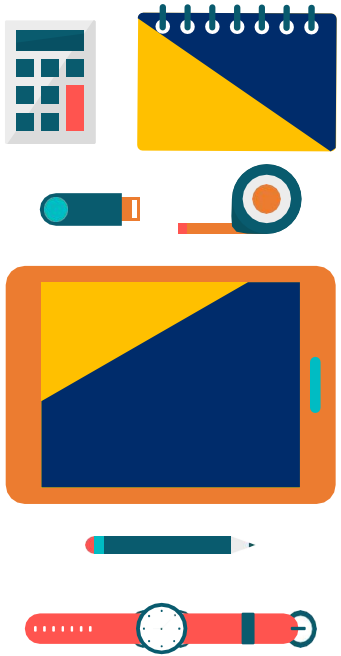
=> Tiết kiệm được tài nguyên, tuy nhiên với tập data lớn thì xác suất gán sai nhãn sẽ tăng.

- Thử với tập data thì gán tự động được 286 rows, có 1 row sai.

	id	title	genre	annotated_genre	match
825	826	kiện nhà mạng vì rằng sóng 5g gây hại tới sức ...	cong-nghe-game	suc-khoe	False

- Visualization:





03

Models

03 Models

- Sử dụng 3 model có sẵn, 1 trong số đó là finetune LLM model.
- Embedding: Sử dụng TF-IDF để vectorize data.
- 2 model Machine Learning sử dụng cách embed trên là Naive Bayes & Support Vector Classifier.
- Finetune 1 LLM model (phoBERT), sử dụng tokenizer của phoBERT.

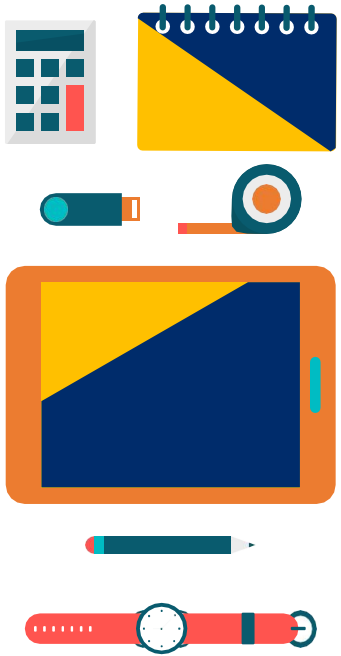
Tuning Parameters

- tfidfVectorizer: có 1 parameter cần tuning là ngram_range. Qua thử nghiệm thì với ngram_range = (1, 2) (tức sử dụng 1-gram và 2-grams) đạt hiệu quả tốt nhất.

- Naive Bayes: cần tuning ‘alpha’: sử dụng StratifiedKFold. Best performance với ‘alpha=0.2’.

- SVC: cần tuning kernel, ‘c’ và ‘gamma’: sử dụng GridSearchCV, trả về thông số tốt nhất với ‘rbf’ kernel, ‘c=10’ và ‘gamma=0.1’.

- Finetuned_phoBERT: train 10 epochs, lr = 2e-5, weight_decay = 0.01.

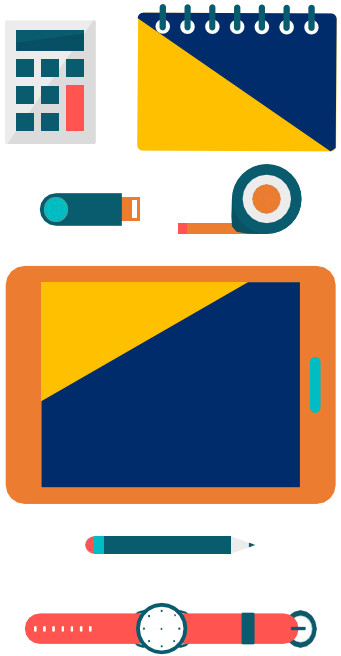


04

Evaluation

Model results table

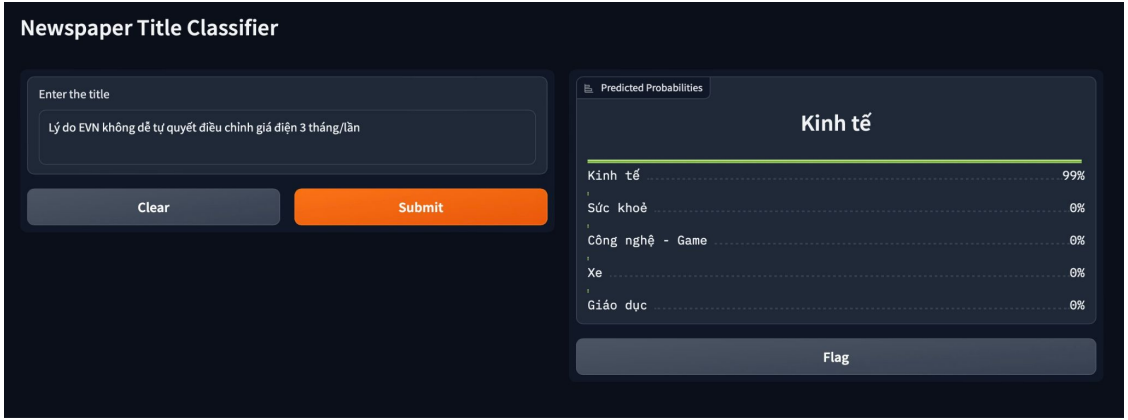
Model	Accuracy	Precision	Recall	F1-score
SVC	88.88%	88.77%	88.88%	88.78%
Naive_Bayes	86.80%	87.07%	86.80%	86.86%
phoBERT_finetuned	**93.60 %**	**93.71%**	**93.60%**	**93.62%**

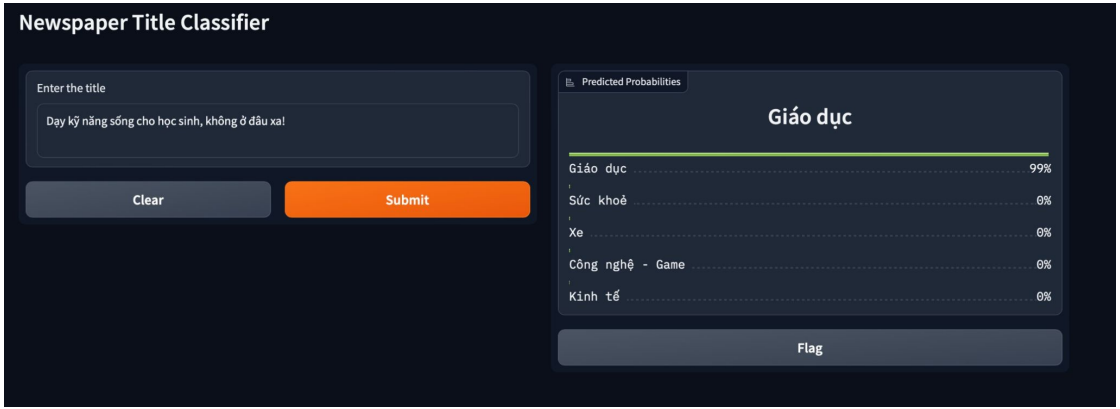


05

Demo & Conclusion

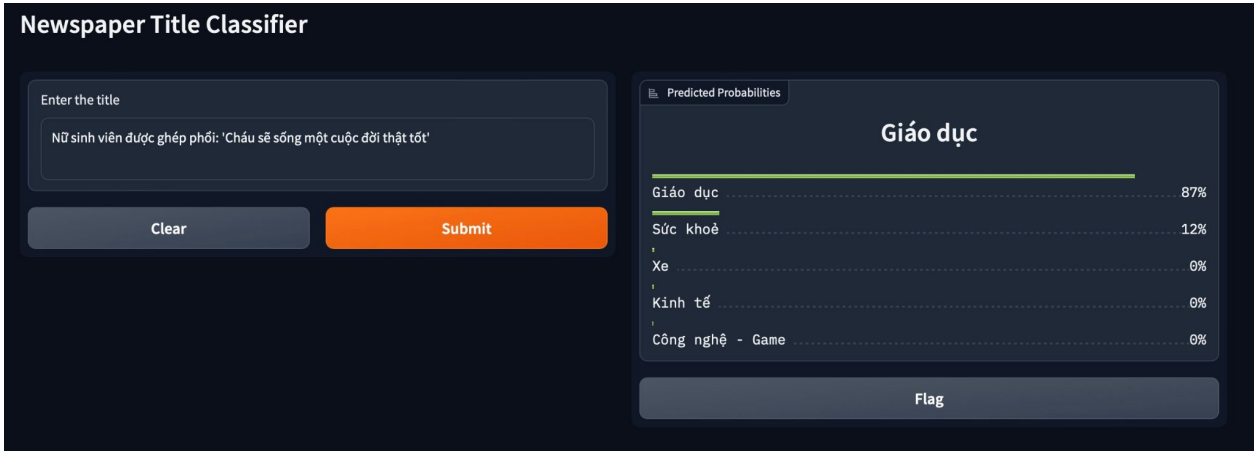
05 Demo & Conclusion





Limitation

- Chưa gặp được nhiều các thách thức của việc gán nhãn.
- Hạn chế về vocab_size vì lượng training data chưa đủ lớn.
- Vẫn chưa thể trả lời đúng 1 số title có độ nhiễu cao.



Future work

- Sử dụng thêm các models LLM khác được pretrained trên tập data tiếng Việt (VD: PhoGPT, Vistral,...)
- Update thêm training data để tăng accuracy của model.
- Sử dụng benchmark data mạnh hơn để đánh giá model.
- Dự đoán các nhãn cụ thể hơn bài báo đó có (VD: kinh tế > ngân hàng).

Contribution

Nhật Tân

- Crawl data
- Gán nhãn
- Làm model SVC
- Làm Demo
- Làm slide

Tiến Dũng

- Crawl data
- Viết Annotation guideline
- Gán nhãn
- Làm model NB
- Làm slide

Việt Anh

- Visualize
- Gán nhãn
- Finetune LLM
- Tuning params
- Evaluation
- Làm slide

Thank you!

Presented by Group 5

Email liên hệ:
22022530@vnu.edu.vn
22022508@vnu.edu.vn
22022644@vnu.edu.vn