ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC
CS114.N21 – MÁY HỌC

ĐỒ ÁN: DỰ ĐOÁN GIỚI TÍNH MỘT NGƯỜI DỰA TRÊN HỌ TÊN TIẾNG VIỆT NGƯỜI ĐÓ

GV hướng dẫn: TS. Lê Đình Duy
ThS. Phạm Nguyễn Trường An

Nhóm sinh viên thực hiện:

1. Nguyễn Trung Kiên 21521024
2. Phạm Quốc Việt 21522792
3. Nguyễn Sỹ Hùng 21522119

Tp.HCM, tháng 07 năm 2023

Tóm tắt những nội dung nhóm đã thực hiện trước khi sửa đổi:

+ Giải quyết bài toán mà nhóm đã nêu ra bằng những gì đã học được từ trên lớp, kèm theo một số lý thuyết tham khảo thêm.

+ Nhóm sử dụng 4 models classification: Logistics Regression, SVM, Gaussian Naïve Bayes và Decision Tree, và xem xét các trích chọn đặc trưng đối với bài toán, đó là trích lọc đặc trưng cho văn bản, kèm theo việc nên bỏ phần "họ" trong họ và tên để xem model có tăng kết quả lên hay không.

+ Nhóm sử dụng giá trị f1-score để đánh giá. Ban đầu nhóm chúng em đọc nckh của thầy cô trong trường thực hiện với bài toán của nhóm, kết quả f1-score đạt max là 96%. Vì thế trước khi thực hiện chúng em đặt mục tiêu f1-score đạt trên 90%. Và kết quả thu được tốt nhất của nhóm đó là f1-score đạt max là 95.13%, sử dụng mô hình SVM và không thực hiện bỏ họ (trong họ và tên)

Cập nhật sau khi vấn đáp:

+ Nội dung trình bày còn sơ sài, không chau chuốt, rõ ràng, mỗi lần xem nội dung phải kéo lên kéo xuống rất bất tiện: Phương án của nhóm đó là đã cập nhật mục lục tự động cho bài báo cáo (Trang 4) và chỉnh sửa lại một số nội dung trong bài báo cáo sao cho hiệu quả với người đọc. Nội dung chương V - Áp dụng các mô hình máy học và trích chọn đặc trưng (Trang 11) và chương VI – Kết luận (Trang 14) được biên soạn lại gần như khác hoàn toàn so với bản báo cáo trước, các chương khác cũng có nhưng không đáng kể.

+ Quá nhiều confusion matrix vào từng model một cách không cần thiết, nhóm quyết định tạo một bảng phương án hai chiều để lưu lại kết quả f1-score cho từng trích chọn đặc trưng văn bản và mô hình machine learning đã chọn (Trang 12)

+ Chưa đưa được những gì đã thực hiện vào ứng dụng thực tiễn. Phương án của nhóm là tạo một API để người khác có thể nhập họ và tên vào để mô hình có thể dự đoán, cách thức thực hiện dựa trên thư viện streamlit (Trang 14) và kèm theo Link colab khởi chạy API sử dụng thư viện streamlit: API.ipynb

+ Chưa so sánh performance của nhóm và của thầy cô trên cùng một bộ dữ liệu. Phương án nhóm đề ra là so sánh trên cùng một bộ dữ liệu và bổ sung vào chương VI - kết luận (Trang 14) và link colab dẫn chứng kết quả chạy của nhóm: Compare.ipynb

+ f1-score ở bản báo cáo cũ là f1-score của giới tính Nam, không phải giá trị trung bình. Phương án giải quyết là sửa đổi, thêm precision và recall để phân tích kết quả (Trang 12)

+ Cập nhật github của nhóm: https://github.com/kiendoo4/final-project

+ Cập nhật link colab final project của nhóm:

https://colab.research.google.com/drive/15RTuAjcwaJR6pPTW21QJI0yr09PWZSmn#scrollTo=un-MHcy_4qF7

+ Link dataset của nhóm: https://github.com/kiendoo4/final-project/blob/main/dataset/Final_dataset.json

+ Kết quả cao nhất cuối cùng sau khi chỉnh sửa là 94.77%, sử dụng Bag of words, mô hình Logistic Regression và lấy đầy đủ họ và tên.

Lời cuối cùng, nhóm chúng em xin được lần nữa cảm ơn thầy **Ths. Phạm Nguyễn Trường An** đã chỉ ra những điểm còn chưa tốt, điểm dở của nhóm, để nhóm có thể khắc phục, sửa đổi và hoàn thành bài báo cáo này. Dù vậy, chắc chắn vẫn sẽ không tránh khỏi được những thiếu sót, mong được thầy đánh giá và góp ý ạ.

Chúng em xin chân thành cảm ơn.

Nhóm HKV.

MỤC LỤC

LỜI NÓI ĐẦU 5

Chương I. Tổng quan đồ án 6

1.1. Mô tả bài toán 6

1.2. Input và output của bài toán 6

1.3. Các thuật toán máy học mà đồ án sử dụng 6

1.4. Các tiêu chí về một mô hình được đánh giá tốt 6

Chương II. Những nghiên cứu về đồ án của các "đại đao cao thủ" 6

2.1. Gender Prediction Based on Vietnamese Names with Machine Learning Techniques 6

2.2. Gender Prediction Based on Chinese Name 7

2.3. GenderAPI 7

Chương III. Xây dựng bộ dữ liệu 7

Chương IV. Các feature extraction mà nhóm thử nghiệm 8

4.1. Lược bỏ họ (Last Name) 8

4.2. Các phương pháp trích chọn đặc trưng cho văn bản 8

4.2.1. PhoBERT 9

4.2.2. Bag of Word (BOW) 9

4.2.3. n-grams 9

4.2.4. Tf-Idf 10

Chương V. Áp dụng các mô hình máy học và trích chọn đặc trưng 11

5.1. Giới thiệu sơ lược các mô hình máy học được sử dụng 11

5.2. Kết quả thu được 12

5.2.1. Họ + Tên đệm + Tên 12

5.2.2. Tên đệm + Tên 12

5.2.3. Phân tích kết quả 12

Chương VI. Kết luận 14

LỜI NÓI ĐẦU

Trong thực tế, phân loại giới tính một người có thể phục vụ cho một số mục đích, công việc. Ví dụ như công việc tiếp thị và quảng cáo sản phẩm, các công ty cần nắm được thông tin giới tính của khách hàng để có thể quảng cáo và đưa ra những sản phẩm thích hợp với đặc điểm giới tính. Quan trọng hơn, đó là việc nghiên cứu về phân phối giới tính trong một phạm vi, khu vực nào đó, nhằm có một cái nhìn trực quan về hiện trạng mất cân bằng giới tính - một giới chiếm tỉ lệ đông hơn hẳn so với phần còn lại. Ngoài ra, với các trang web yêu cầu nhập thông tin user, sau khi người dùng nhập họ và tên của người đó thì ứng dụng đó có thể thêm chức năng dự đoán giới tính hiển thị trên màn hình, tất nhiên là vẫn có thể chỉnh sửa nhưng mục đích giúp người dùng tiết kiệm thời gian. Hay cũng có thể chỉ là một vấn đề đơn giản như việc bạn có một dataset rất lớn với rất nhiều cột, ví dụ như cột họ và tên, giới tính, độ tuổi, địa chỉ nhà, email, ... để phục vụ cho mục đích gì đó, bỗng nhiên vì một sự cố nào đó mà cột giới tính bị xóa đi và ta cần phải tìm lại. Nếu tập dataset có số lượng đối tượng lớn, thì việc đi thu thập lại từ đầu sẽ tốn rất nhiều thời gian và chi phí. Vậy nên, nếu như có được một mô hình máy học đủ tin cậy, ta có thể sử dụng nó để dự đoán giới tính mà không phải đi tìm từng người để biết giới tính của họ là gì.

Tóm lại, các ví dụ kể trên muốn ám chỉ rằng chúng ta hoàn toàn có thể làm những công việc đó một cách thủ công, tuy nhiên quy mô thực hiện có thể là quá nhiều và quá lớn để làm hết, và đồng thời, là sự cần thiết của đồ án vào các vấn đề thực tế. Một số những dự án trong thực tế như Gender-API đã được thực hiện và người dùng muốn thực hiện nhiều requests sẽ cần phải trả phí cũng đủ để thấy "cung - cầu" mà bài toán hướng tới trong thực tế là không nhỏ. Những lý đó chính là **lý do** thôi thúc chúng em thực hiện đồ án này.

Chúng em xin chân thành gửi lời cảm ơn đến thầy, **ThS. Phạm Nguyễn Trường An** đã nhiệt tình giảng dạy môn **Máy học – CS114** và truyền đạt những kiến thức, kinh nghiệm vô cùng quý giá cho chúng em trong thời gian học tập trên lớp. Chúng em đã được trang bị những kiến thức quan trọng, các kỹ năng thực tế để có thể hoàn thành đồ án này. Chúng em hi vọng rằng những gì mà mình đã thực hiện có thể phần nào cống hiến cho khoa học và cho xã hội.

Dù vậy, trong quá trình làm đồ án cuối kỳ vẫn khó tránh khỏi những sai sót, chúng em rất mong nhận được sự góp ý của thầy để đồ án có thể hoàn thiện hơn.

Thành phố Hồ Chí Minh, tháng 7 năm 2023.

Chương I. Tổng quan đồ án

1.1. Mô tả bài toán

Đề tài dự đoán giới tính một người dựa vào họ tên là đề tài hướng đến việc sử dụng họ và tên của một người để có thể phân loại người đó là nam hay nữ. Trong Machine Learning, đây là dạng bài toán có tên gọi Binary Classification, trong đó bài toán chỉ có hai classes, cũng là hai giới tính cần được phân loại.

1.2. Input và output của bài toán

- Input: Dataset bao gồm họ và tên và giới tính tương ứng, họ tên người cần dự đoán giới tính.
- Output: Giới tính được dự đoán của người đó.

1.3. Các thuật toán máy học mà đồ án sử dụng

Trong bài báo cáo, nhóm sẽ sử dụng 4 mô hình classification, đó là: Logistics Regression, SVM, Gaussian Naive Bayes và Decision Tree. Sau khi qua các bước xử lý dữ liệu và trích chọn đặc trưng, data được xử lý sẽ đưa vào các mô hình classification và sau cùng đánh giá từng mô hình bằng confusion matrix và giá trị của các metrics: accuracy, precision, recall và f1-score.

1.4. Các tiêu chí về một mô hình được đánh giá tốt

Tiêu chí mà nhóm sử dụng để đánh giá một mô hình có thể xem là tốt đó là:

- f1-score: f1- score là giá trị trung bình điều hòa của precision và recall. Hiểu một cách đơn giản, f1-score cân bằng hai giá trị precision và recall. Đôi khi trong một số trường hợp, tập dữ liệu quá nghiêng về trường hợp positive, giá trị precision sẽ cao nhưng recall thấp, và ngược lại. Hiện tượng này thường được đề cập với tên gọi precision-recall trade-off. Do đó, để tránh việc này, f1-score phải cao thì mới tương ứng với precision cao và recall cao. Qua những nghiên cứu trước đây đã được thực hiện (sẽ được trình bày cụ thể hơn trong chương II), nhóm hi vọng f1-score của nhóm đạt trên 90% khi chưa thực hiện.

Chương II. Những nghiên cứu về đồ án của các "đại đao cao thủ"

2.1. Gender Prediction Based on Vietnamese Names with Machine Learning Techniques

Là bài báo khoa học của nhóm tác giả: **Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, Anh Gia-Tuan Nguyen**. Kết quả của các thầy cô trong trường mình có F1-score rất cao với 96% khi sử dụng Long short-term memory (LSTM) model. Dataset của nhóm thực hiện bao gồm 26850 họ và tên và giới tính kèm theo. Thách thức của nhóm đó là vấn đề về dự đoán sai với phần lớn họ và tên nữ. Ngoài ra, nhóm

cũng hướng đến việc cải tiến đó là mở rộng dataset về tính đa dạng của họ và tên tiếng Việt và đánh giá sự hiệu quả các mô hình transfer learning như BERT và những mô hình học sâu khác.

2.2. Gender Prediction Based on Chinese Name

Là bài báo khoa học của nhóm tác giả: **Jizheng Jia, Qiyang Zhao**. Họ sử dụng BERT model để chuyển hóa chữ giản thế và chữ Hán, sử dụng các mô hình classification kết hợp fastText - một thư viện để học cách nhúng từ và phân loại văn bản được tạo bởi phòng thí nghiệm Nghiên cứu AI của Facebook. Kết quả, phương pháp của nhóm tác giả có 93.45% test dự đoán đúng.

2.3. GenderAPI

Sáng lập vào năm 2014, bởi **Markus Per**. Đây là một công cụ mà ta có thể search trên mạng, giúp dự đoán giới tính qua tên của một ai đó. Tính đến thời điểm nhóm thực hiện đồ án, số lượng tên đã lên đến 6,084,389, bao gồm 190 quốc gia. Nguồn dữ liệu khổng lồ này cũng đảm bảo cho việc dự đoán chính xác hơn. Tuy vậy thì nhược điểm của nó cũng là nguồn từ điển có giới hạn và không phải một open-source.

=> Nhận xét từ các bài báo: ngoài 3 công trình đã nêu, vẫn còn những công trình đề tài nghiên cứu khác, về các ngôn ngữ khác mà nhóm không thể đưa vào hết. Sau khi chốt đề tài thực hiện, xem thử thế giới xung quanh đã làm được những gì và tóm lược lại về những nghiên cứu của các "tiền bối" mà thực hiện, các thành viên trong nhóm không khỏi trầm trồ thán phục, chẳng hạn là f1-score 96% với công trình của các thầy cô trường mình đã thực hiện. Nhóm tác giả Trung Quốc có độ chính xác 93.45% cũng là rất "đáng sợ" khi mà chữ Trung Quốc, dù là giản thể, nhưng cũng là phức tạp với mặt chữ mang ý nghĩa và chi tiết khác biệt nhau, hay GenderAPI đã đưa đề tài dự đoán giới tính qua tên từ tận năm 2014 và thu về được lợi nhuận. Những điều đó cũng là thử thách không nhỏ với nhóm chúng em khi thực hiện đồ án này.

Chương III. Xây dựng bộ dữ liệu

- Nguồn dữ liệu được lấy từ những file excel đã được public của các trường đại học, là danh sách sinh viên của một trường, được các thành viên nhóm tổng hợp lại, xử lý và bỏ đi các feature không cần thiết. Hơn nữa, dataset của nhóm được merge 10000 họ và tên và giới tính tương ứng, lấy từ file excel danh sách của bài báo khoa học phần 2.1. Cuối cùng gọp tất cả lại và convert thành file json. Việc trích các nguồn đã sử dụng được để dưới phần tài liệu tham khảo.

- Dataset ban đầu chứa cả tên một số sinh viên không phải người Việt Nam đã được lược bỏ đi vì quy mô của đề tài, tên của một số người dân tộc được giữ lại để tăng sự đa dạng cho dataset, đây cũng là một tiêu chí mà nhóm đặt ra với dataset của mình.
- Dữ liệu của nhóm chúng em bao gồm 28801 họ và tên, kèm theo đó là giới tính của người đó, nam hoặc nữ. 2 thông tin chính trong file json là full_name (họ và tên) và gender (giới tính) có giá trị được quy ước là 1 với nam giới, và 0 đối với nữ giới.
- Data chứa 16613 người là nam và 12188 người là nữ.
- Mô tả thông số về sự đa dạng của bộ dữ liệu: Dataset sau khi lọc đi những họ và tên trùng thì tập họ và tên trùng số lượng còn lại là 19431. Gần 20000 họ và tên khác nhau, với nhóm là đã đạt được sự "đa dạng".
- Các trường hợp khó xử lý và hướng giải quyết:
+ Ngoài các tên gọi thông thường, bởi phần lớn dân số nước ta là dân tộc Kinh, thì cũng có một số các họ và tên trông khá lạ lẫm đến từ các dân tộc khác. Hướng giải quyết của nhóm đó là vẫn sẽ giữ nguyên để đảm bảo sự đa dạng của tập dataset.
+ Tập họ và tên bị trùng, nhóm xử lý bằng cách loại bỏ đi các họ và tên bị trùng

---

Chương IV. Các feature extraction mà nhóm thử nghiệm

4.1. Lược bỏ họ (Last Name)

- Nhận xét: Thông thường, trong họ và tên của người Việt chúng ta, phần "họ" thường là tên các triều đại trong lịch sử dân tộc Việt Nam, mang ý nghĩa rằng chúng ta thuộc về một gia tộc hay dòng dõi nào đó. Các họ phổ biến ở Việt Nam thường là: Nguyễn, Trần, Lê, Lý, Phạm … Trích Wikipedia, khoảng 31.5% dân số Việt Nam mang họ Nguyễn, 10.9% dân số mang họ Trần. Qua thống kê trên, ta thấy rằng sự đa dạng về họ của người Việt Nam là rất chênh lệch. Hơn nữa, như đã trình bày, ý nghĩa của "họ" trong họ và tên của người Việt Nam cũng không mang nhiều ý nghĩa về giới tính.
- Một ví dụ đơn giản, lấy họ và tên của một người Việt Nam, thì xu hướng của chúng ta sẽ là nhìn vào tên đệm và tên gọi của người đó để đoán xem là họ mang giới tính gì. Ví dụ: Trung Kiên thì có thể là nam, Ngọc Hoa thì có thể là nữ…
- Tóm lại, nhóm sẽ đánh giá các mô hình classification khi có họ trong tên và cả khi không có, sau cùng là đánh giá kết quả. Phần này sẽ được trình bày kỹ hơn ở phần V.

4.2. Các phương pháp trích chọn đặc trưng cho văn bản

Dữ liệu văn bản có thể tồn tại ở nhiều dạng khác nhau như chữ cái thường, chữ cái hoa, dấu câu, các kí tự đặc biệt…. Các ngôn ngữ khác nhau cũng có mẫu kí tự khác nhau và cấu trúc ngữ pháp khác nhau.

Vấn đề chính của dữ liệu dạng văn bản đó là làm thể nào để mã hoá được kí tự về dạng số? Kĩ thuật tokenization sẽ giúp ta thực hiện điều này. tokenization là việc chúng ta chia văn bản theo đơn vị nhỏ nhất và xây dựng một từ điển đánh dấu index cho những đơn vị này. Dưới đây là các phương pháp mà nhóm sẽ sử dụng

4.2.1. PhoBERT

PhoBERT: là một pre-trained language models dành riêng cho tiếng Việt. Các kết quả thực nghiệm cho thấy PhoBERT vượt trội hơn hẳn so với các mô hình XLM-R - một một model đa ngữ được sử dụng rộng rãi, và còn là mô hình tốt nhất cho các tác vụ dành riêng cho xử lý text tiếng Việt

4.2.2. Bag of Word (BOW)

Bag of Words (BoW) là một phương pháp tiền xử lý dữ liệu trong Machine Learning, đặc biệt trong các bài toán xử lý ngôn ngữ tự nhiên (NLP). BoW được sử dụng để biểu diễn và đại diện cho văn bản dưới dạng các vectơ số học.

Ý tưởng cơ bản của BoW là chuyển đổi các văn bản thành tập hợp các từ và đếm số lần xuất hiện của từng từ trong văn bản. Việc này không quan tâm đến thứ tự hay ngữ nghĩa của từng từ, chỉ quan tâm đến tần suất xuất hiện của chúng. Một lần biểu diễn BoW được xây dựng, các văn bản ban đầu có thể được so sánh và phân loại bằng các phương pháp học máy khác nhau như học có giám sát (supervised learning) hoặc học không giám sát (unsupervised learning).

Nó bao gồm hai điều:

- Một từ điển các từ đã biết.
- Một phép đo về sự xuất hiện của các từ đã biết.

Nó được gọi là "túi" các từ vì bất kỳ thông tin nào về thứ tự hoặc cấu trúc của các từ trong văn bản đều bị loại bỏ. Mô hình chỉ quan tâm đến việc các từ đã biết xuất hiện trong văn bản, không phải vị trí của chúng trong văn bản.

4.2.3. n-grams

n-grams: là một extension của phương pháp bag of words. Nó được sử dụng nhiều trong các tác vụ NLP (xử lý ngôn ngữ tự nhiên) với các việc kiểm tra sửa lỗi chính tả, nhận dạng ngôn ngữ, khai thác từ vựng, dự đoán từ tiếp theo.

Ý tưởng cơ bản của n-grams đó là xem xét các cụm từ chứ n từ liên tiếp để hiểu ngữ cảnh và tương quan giữa các từ trong văn bản. Điều này giúp ta thu thập thông tin về cấu trúc ngôn ngữ, biểu đạt ý nghĩa và tính chất ngữ pháp của văn bản.

Ví dụ trường hợp như sau: ta có một câu như sau “Bầu trời hôm nay đẹp thật đấy.”, ta chia câu này thành các phần như sau:

1-grams (unigram): Bầu, trời, hôm, nay, đẹp, thật, đấy.

2-grams (bigram): Bầu trời, trời hôm, hôm nay, nay đẹp, đẹp thật, thật đấy.
3-grams (trigram): Bầu trời hôm, trời hôm nay, hôm nay đẹp, nay đẹp thật, đẹp thật đấy.

4.2.4. Tf-Idf

Term frequency-inverse document frequency (Tf-Idf) là một trong những phương pháp được sử dụng rộng rãi hiện nay trong lĩnh vực xử lý dữ liệu văn bản. Tf-Idf thể hiện mức độ quan trọng của một từ trong văn bản mà văn bản đang xét nằm trong một tập hợp các văn bản. Tf-Idf được phân làm 2 phần: TF(term frequency) với IDF(Inverse Document Frequency).
+ TF là tần suất một từ xuất hiện trong một văn bản. Vì có những văn bản với độ dài khác nhau và số lần xuất hiện của từ có thể nhiều hơn, thế nên tf được tính bằng số lần xuất hiện của từ chia cho tổng số từ trong văn bản đó.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}},$$

Trong đó: $f_{t,d}$ là số lần xuất hiện của từ trong văn bản, $\sum_{t' \in d} f_{t',d}$ là tổng số từ trong cùng một văn bản.
+ IDF dùng để ước lượng mức độ quan trọng của từ đó như thế nào. Ví dụ như có các từ xuất hiện nhiều nhưng mức độ quan trọng của nó lại không được cao như các từ nối (và, nhưng,..), giới từ (ở, trong, trên,...). Thế nên ta cần phải dùng IDF để bỏ bớt đi những từ đó với công thức như sau:

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

.
Trong đó: N là số lượng các văn bản trong tập D, $|\{d \in D : t \in d\}|$ là số lượng văn bản có chứa từ t trong đó.
Tổng kết lại giá trị TF-IDF được tính như sau:

TF-IDF = TF * IDF

Giá trị TF-IDF này có ý nghĩa là: nếu như giá trị này cao đồng nghĩa với việc từ đó có khả năng rất cao là từ đặc biệt trong tài liệu, ít xuất hiện trong những tập tài liệu khác. Ngược lại nếu giá trị này thấp chứng tỏ từ đó có độ quan trọng thấp (thường là các từ thông thường), dễ tìm thấy trong những tập tài liệu khác, những từ này bị loại bỏ và không cần thiết cho việc xử lý dữ liệu.

Chương V. Áp dụng các mô hình máy học và trích chọn đặc trưng

- Dataset được chia 80% cho training mô hình máy học và 20% cho việc testing.

5.1. Giới thiệu sơ lược các mô hình máy học được sử dụng

- Logistic Regression: là một thuật toán phân loại thường được sử dụng khi các nhóm class là rời rạc. Ví dụ cơ bản nhất là việc phân loại bệnh nhân có bị ung thư hay không, hay mail gửi đến có phải là mail spam không. Thuật toán này khá giống với Linear Regression, tuy nhiên tập class là rời rạc, và dữ liệu đầu vào cho mô hình cũng được ánh xạ đến một hàm sigmoid f(s) như sau để tập dữ liệu đưa vào Logistic Regression nằm trong khoảng (0, 1).

$$f(s) = \frac{1}{1 + e^{-s}}$$

- Gaussian Naive Bayes: Mô hình phân loại Naive Bayes xác định khả năng một text dữ liệu x rơi vào class c, bằng cách tìm max(p(c|x)). Với đầu vào c và các class x_i, nếu tìm được x_i sao cho p(c|x_i) đạt max thì x_i chính là class mà c rơi vào. Trong bài toán binary classification, áp dụng quy tắc Bayes, ta sẽ có c = $\arg(\max_{c=\{0,1\}} p(c|x)) = \arg(\max_{c=\{0,1\}} p(x|c).p(c))$. Trong đó p(c) có thể được xác định như là tần suất xuất hiện của class c trong training set. Với phân phối Gaussian Naive Bayes, mỗi chiều dữ liệu i và một class c, x_i tuân theo một phân phối chuẩn có kỳ vọng μ_{ci} và phương sai $(\sigma_{ci})^2$, p(x_i|c) được tính như sau:

$$p(x_i|c) = p(x_i|\mu_{ci}, \sigma_{ci}^2) = \frac{1}{\sqrt{2\pi\sigma_{ci}^2}} \exp\left(-\frac{(x_i - \mu_{ci})^2}{2\sigma_{ci}^2}\right)$$

- SVM: Trong bài toán này, các điểm dữ liệu đầu vào là rời rạc trên một không gian nhiều chiều. Về lý thuyết, SVM sẽ tìm một siêu mặt phẳng để phân chia các điểm dữ liệu đầu vào thành 2 vùng không gian phân biệt, sao cho tổng khoảng cách giữa các điểm dữ liệu đầu vào tới siêu mặt phẳng là nhỏ nhất. Với một đoạn text dữ liệu đã được chuyển hóa thành một dạng số hóa nào đó, ví dụ như numpy array, ta có thể sử dụng thuật toán SVM sau khi đã được training cho việc phân loại giới tính.

- Decision Tree: là một thuật toán có thể giải quyết cả bài toán dạng regression và classification. Ý tưởng chính của thuật toán trong đồ án này là đưa dữ liệu vào mô hình Decision Tree cho việc training nhằm tạo ra một cây quyết định, sau đó đi theo các điều kiện từ node gốc dần xuống node lá để phân định xem giới tính người đó là nam hay nữ.

5.2. Kết quả thu được

5.2.1. Họ + Tên đệm + Tên

- Dưới đây là kết quả f1-score tương ứng với mô hình machine learning và trích chọn đặc trưng tương ứng, lấy toàn bộ họ và tên, với class 0(Nữ) và 1(Nam)

	Logistic Regression			**SVM**			**GaussianNB**			**Decision Tree**		
	0	**1**	**Avg**	**0**	**1**	**Avg**	**0**	**1**	**Avg**	**0**	**1**	**Avg**
PhoBERT	92.62	94.66	93.64	92.73	94.86	93.80	88.15	92.08	90.11	79.14	84.16	81.65
Bag of words	93.91	95.62	**94.77**	93.76	95.51	94.64	68.95	53.13	61.04	89.42	92.05	90.73
n-grams	84.20	90.41	87.30	84.84	90.64	87.74	80.93	80.43	80.68	84.44	90.01	87.22
TF-IDF	93.11	95.09	94.10	93.65	95.46	94.55	68.17	50.17	59.17	88.39	91.25	89.82

Bảng I: f1-score của các mô hình máy học và trích chọn đặc trưng tương ứng cho văn bản, sử dụng họ + tên đệm + tên

5.2.2. Tên đệm + Tên

- Dưới đây là kết quả f1-score tương ứng với mô hình machine learning và trích chọn đặc trưng tương ứng, lấy tên đệm và tên, với class 0(Nữ) và 1(Nam)

	Logistic Regression			**SVM**			**GaussianNB**			**Decision Tree**		
	0	**1**	**Avg**	**0**	**1**	**Avg**	**0**	**1**	**Avg**	**0**	**1**	**Avg**
PhoBERT	88.52	91.80	90.16	89.30	92.46	90.88	81.05	86.27	83.66	83.35	87.22	85.29
Bag of words	91.14	93.63	**92.39**	90.95	93.46	92.21	69.85	56.07	62.96	86.87	89.97	88.42
n-grams	76.30	86.58	81.44	76.30	86.58	81.44	73.34	65.45	69.39	76.60	86.45	81.53
TF-IDF	91.05	93.60	92.33	91.11	93.61	92.36	69.49	54.71	62.10	88.40	91.32	89.86

Bảng II: f1-score của các mô hình máy học và trích chọn đặc trưng tương ứng cho văn bản, sử dụng tên đệm + tên

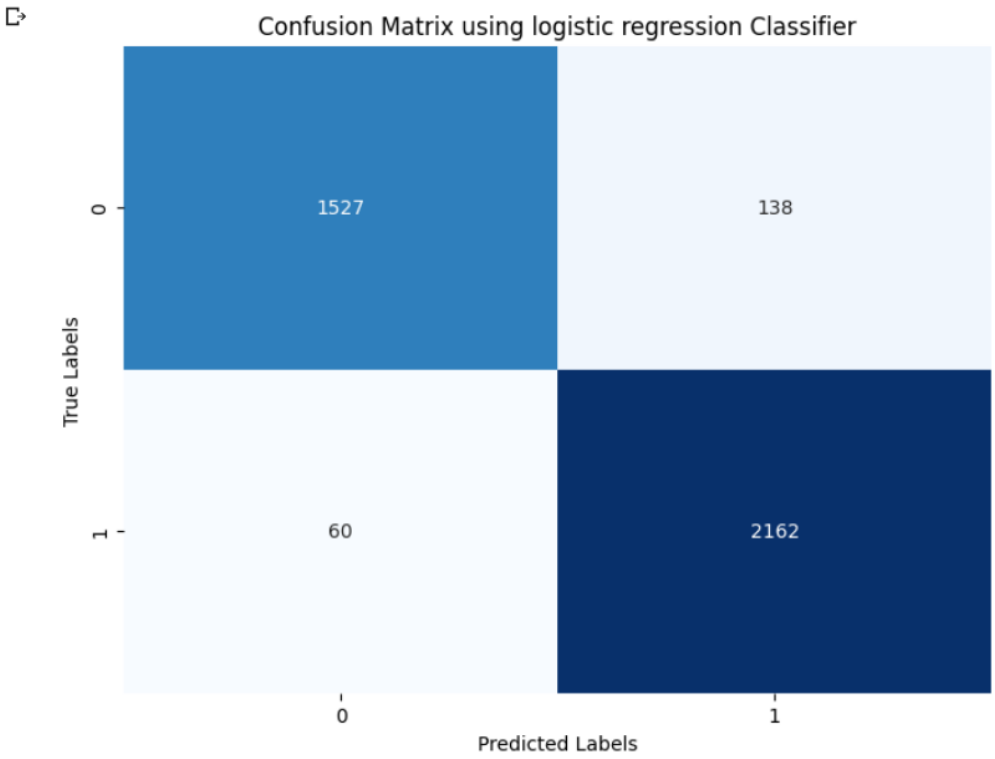
5.2.3. Phân tích kết quả

- Sau khi chạy Colab và thu được các kết quả từ bảng I và II, kết quả f1-score cao nhất mà nhóm thu được là **94.77%** khi sử dụng Bag of words và Logistic Regression, lấy toàn bộ họ và tên. Các thông số cụ thể hơn như bên dưới

	Precision	**Recall**	**F1-score**	**Support**
0	0.96219	0.91712	0.93911	1665
1	0.94000	0.97300	0.95621	2222
Accuracy			0.94906	3887
Precision score of using Logistic Regression:			0.9510964083175804	
Recall score of using Logistic Regression:			0.945057208423545	
F1-score of using Logistic Regression:			0.9476642278585545	

Bảng III: Kết quả cụ thể Bag of words + Logistic Regression, lấy toàn bộ họ và tên

- F1-score max mà nhóm có được xấp xỉ 95%, điều này vượt mong đợi của nhóm. Tuy nhiên, vấn đề đối với 5% còn lại là gì? Ta có thể nhìn vào confusion matrix dưới đây để xem chi tiết performance kết quả cao nhất đạt được:



Hình I: Confusion matrix, khi sử dụng Bag of words
+ Logistic Regression, lấy toàn bộ họ và tên

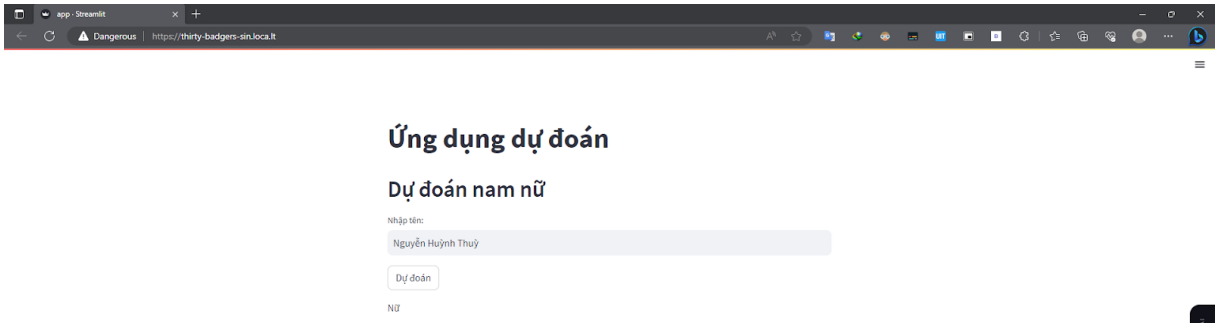
- Qua confusion matrix, ta nhận thấy số lượng các họ và tên của nữ giới bị dự đoán là nam là 138, còn họ và tên nam giới bị dự đoán là nữ là 60. Vậy thì, họ và tên của nữ giới có xu hướng bị đoán nhầm cao hơn, điều này cũng xảy ra với bài báo khoa học của các thầy cô trong trường. Lý do là vì một số tên của nữ giới dễ bị hiểu nhầm là tên của nam.
- Một số tên hay bị nhầm lẫn được thể hiện trong bảng IV.

Họ và tên	Giới tính dự đoán	Giới tính thật sự
Nguyễn Hồng Khánh	Nữ	Nam
Huỳnh Phạm Hồng Thủy	Nữ	Nam
Cao Thị Anh Tuyết	Nữ	Nam
Trần Lâm Hân	Nữ	Nam
Hồ Nhựt Tiến	Nam	Nữ
Trần Nguyễn Thanh Bình	Nam	Nữ
Nguyễn Thái Phi	Nam	Nữ
Võ Kim Điền	Nam	Nữ

Bảng IV: Một số họ và tên dễ bị nhầm lẫn giới tính trong dataset

Chương VI. Kết luận

- Trước hết, kết quả f1-score cao nhất mà nhóm thu được là 94.77% khi sử dụng Bag of words và Logistic Regression, lấy cả họ và tên.
- So sánh kết quả của nhóm với kết quả của thầy cô, dựa trên cùng một dataset của thầy cô (khác với dataset mà nhóm đã train, test), chúng em đã so sánh, đối chiếu kết quả (f1-score) với thầy cô qua kết quả tốt nhất của cả hai, kết quả thu được đó là f1-score cao nhất là 95.34%, sử dụng Bag of Word, mô hình Logistic Regression, lấy cả họ, tên đệm và tên. Kết quả cao nhất của các thầy cô là 95.89%, sử dụng mô hình LSTM. Hai kết quả này không quá chênh lệch, tuy nhiên các thầy cô vẫn là nhỉnh hơn chúng em một chút. Dẫn chứng: Compare.ipynb
- Ngoài ra, qua đồ án này, chúng em đã tạo một api cho việc phân loại giới tính. Mã nguồn của code: API.ipynb - Colaboratory
- Api này được xây dựng dựa trên mô hình Bag of Words (BOW) và được huấn luyện bằng mô hình máy học Logistic Regression vì qua quá trình thử nghiệm ở phía trên cho thấy mô hình này là tốt nhất cho việc phân loại giới tính dựa trên họ và tên. Ta cần thu thập dữ liệu (ví dụ ở đây là dữ liệu ban đầu của nhóm), phân tích dữ liệu qua mô hình Bag Of Words (BOW), huấn luyện cho máy học dựa trên mô hình Logistic Regression. Tiếp theo ta cần cài đặt streamlit và localtunnel. Streamlit là một framework mã nguồn mở bằng ngôn ngữ Python, rất thích hợp cho việc tạo một trang web cho máy học (machine learning) trong thời gian ngắn. Localtunnel là một công cụ giúp ta dễ dàng chia sẻ trang web của mình với người khác trên mạng cục bộ mà không ảnh hưởng gì tới DNS và các thiết lập tường lửa. Để tạo app bằng streamlit, ta tạo một tệp để lưu mô hình đã train, sau đó ta load tệp này. Tiếp theo ta tạo những thứ đơn giản như tiêu đề, một khoảng trống để ta nhập input họ và tên người cần dự đoán, một button để thực hiện thao tác dự đoán và một văn bản để hiển thị giới tính mà máy dự đoán rồi ta chạy ứng dụng. Ta lấy địa chỉ IP và với localtunnel, nó tạo ra một trang web dẫn tới trang web dự đoán giới tính. Dựa theo địa chỉ IP ta đã lấy, ta nhập vào trong trang web. Giao diện đơn giản như hình 2.



Hình 2: Trang web ứng dụng dự đoán giới tính giới tính dựa trên họ và tên

Tài liệu tham khảo

[1] PhoBERT: https://github.com/VinAIResearch/PhoBERT#transformers

[2] Các bài báo khoa học được trích trong bài

Gender Prediction Based on Vietnamese Names with Machine Learning Techniques: https://arxiv.org/ftp/arxiv/papers/2010/2010.10852.pdf

Gender Prediction Based on Chinese Name: http://tcci.ccf.org.cn/conference/2019/papers/SW4.pdf

[3] Trích chọn đặc trưng cho văn bản: https://viblo.asia/p/machine-learning-trich-xuat-dac-trung-van-ban-part-1-oOVlYqzzl8W

[4] Các phương pháp đánh giá một hệ thống phân lớp: https://machinelearningcoban.com/2017/08/31/evaluation/#-confusion-matrix

[5] Tham khảo các xây dựng dữ liệu bằng Bag of Words

How to Develop a Deep Learning Bag-of-Words Model for Sentiment Analysis (Text Classification) - MachineLearningMastery.com

[6] TfidfVectorizer sklearn: sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn 1.3.0 documentation

[7] Tf-idf Wikipedia: tf–idf - Wikipedia

[8] Tf-idf Capital One: Understanding TF-IDF for Machine Learning | Capital One

[9] n-gram Wikipedia: n-gram - Wikipedia

[10] CountVectorizer sklearn: sklearn.feature_extraction.text.CountVectorizer — scikit-learn 1.3.0 documentation

[11] Tham khảo về GenderAPI: https://gender-api.com/

[12] Nguồn database mà nhóm sử dụng:

+ Database từ nhóm các thầy cô trong trường UIT: https://github.com/JkUndead/UIT-ViNames-Dataset?fbclid=IwAR1bXso9_9YbmVbPvCm8heebNTiXpI6Rhx7tMQQh3qPiJcyUCBbi6QXCMJw

+ Danh sách sinh viên tốt nghiệp học kỳ II, trường đại học Thủy Lợi, năm học 2015-2016 và 2020-2021: http://www.tlu.edu.vn/sinh-vien-tot-nghiep/danh-sach-sinh-vien-tot-nghiep-7906

+ Danh sách sinh viên tốt nghiệp kỳ 2020.2B, viện Công nghệ Thông tin và Truyền thông:https://docs.google.com/spreadsheets/d/1YtsBTLlt7n8z0OKLuh9dBPvRV9OCrmz4/edit?fbclid=IwAR2P77Xh6t6ytxDpeFieMtqZQRTokRJaC-KTMMX_Oe7WbN79WM_7uBnJ174#gid=1167259155

+ Danh sách sinh viên đạt học bổng khuyến khích học tập: https://ctsv.ntt.edu.vn/wp-content/uploads/2023/06/16-06-2023_16.48.53dssv-chinh-thuc-dat-hoc-bong-khuyen-

khich-21-22-dang-web.pdf?fbclid=IwAR3U-FDOHpi-uWEfoXGwIYsOVQ83saIlHpNULHojTY8pgxPRy43BpDd0IAY
+ Danh sách sinh viên các khoa, các năm của trường đại học Y: https://huemed-univ.edu.vn/modules.php?name=65nam&file=alumni&fbclid=IwAR2uCzp5TKksp8oL7WBAbSmOovOjReAQfJhxFEq7tTRX3wUG3kIWZ6F7cLA
[13] Logistic Regression: https://machinelearningcoban.com/2017/01/27/logisticregression/
[14] SVM: https://machinelearningcoban.com/2017/04/09/smv/
[15] Gaussian Naive Bayes: https://machinelearningcoban.com/2017/08/08/nbc/
[16] Decision Tree: https://machinelearningcoban.com/tabml_book/ch_model/decision_tree.html
[17] Thư viện Streamlit: https://streamlit.io/
[18] Localtunnel: https://theboroer.github.io/localtunnel-www/