**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**---------------------------------------**



**TRƯƠNG CÔNG HẢI**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI**

**DỰA TRÊN NỘI DUNG BÀI VIẾT**

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

HÀ NỘI – 2016

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**---------------------------------------**



**TRƯƠNG CÔNG HẢI**

**DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI**

**DỰA TRÊN NỘI DUNG BÀI VIẾT**

|  |  |  |
| --- | --- | --- |
| **CHUYÊN NGÀNH** | **:** | **KHOA HỌC MÁY TÍNH** |
| **MÃ SỐ** | **:** | **60.48.01.01** |

**LUẬN VĂN THẠC SĨ KỸ THUẬT**

***(Theo định hướng ứng dụng)***

NGƯỜI HƯỚNG DẪN KHOA HỌC

**PGS.TS. TỪ MINH PHƯƠNG**

HÀ NỘI - 2016

LỜI CAM ĐOAN

Luận văn này là thành quả của quá trình học tập nghiên cứu của em cùng sự giúp đỡ, khuyến khích của các quý thầy cô sau 2 năm em theo học chương trình đào tạo Thạc sĩ, chuyên ngành Khoa học máy tính của trường Học viện Công nghệ Bưu chính Viễn thông.

Em cam đoan đây là công trình nghiên cứu của riêng em. Nội dung của luận văn có tham khảo và sử dụng một số thông tin, tài liệu từ các nguồn sách, tạp chí được liệt kê trong danh mục các tài liệu tham khảo và được trích dẫn hợp pháp.

|  |  |
| --- | --- |
|  | Tác giả  *(Ký và ghi rõ họ tên)*  **Trương Công Hải** |

LỜI CÁM ƠN

Em xin gửi lời cảm ơn và tri ân tới các thầy cô giáo, cán bộ của Học viện Công nghệ Bưu chính Viễn thông đã giúp đỡ, tạo điều kiện tốt cho em trong quá trình học tập và nghiên cứu chương trình Thạc sĩ.

Em xin gửi lời cảm ơn sâu sắc tới **PGS.TS. Từ Minh Phương** đã tận tình hướng dẫn, giúp đỡ và động viên em để hoàn thành tốt nhất Luận văn “DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI DỰA TRÊN NỘI DUNG BÀI VIẾT“.

Do vốn kiến thức lý luận và kinh nghiệm thực tiễn còn ít nên luận văn không tránh khỏi những thiếu sót nhất định. Em xin trân trọng tiếp thu các ý kiến của các thầy, cô để luận văn được hoàn thiện

Trân trọng cám ơn.

|  |  |
| --- | --- |
|  | Tác giả  *(Ký và ghi rõ họ tên)*  **Trương Công Hải** |

MỤC LỤC

[MỤC LỤC iii](#_Toc484733162)

[DANH MỤC TỪ VIẾT TẮT vi](#_Toc484733163)

[DANH MỤC CÁC BẢNG BIỂU viii](#_Toc484733164)

[DANH MỤC CÁC HÌNH VẼ ix](#_Toc484733165)

[MỞ ĐẦU 1](#_Toc484733166)

[Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH 3](#_Toc484733167)

[1.1. Giới thiệu bài toán dự đoán giới tính. 3](#_Toc484733168)

[1.2. Các phương pháp dự đoán giới tính 3](#_Toc484733169)

[1.3. Các phương pháp dự đoán giới tính dựa trên các bài biết của người dùng 4](#_Toc484733170)

[*1.3.1.* *Dự đoán giới tính sử dụng bài viết từ blog* 4](#_Toc484733171)

[*1.3.2.* *Dự đoán giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy* 5](#_Toc484733172)

[1.4. Kết luận chương 7](#_Toc484733173)

[Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI 8](#_Toc484733174)

[2.1. Các đặc trưng sử dụng 8](#_Toc484733175)

[*2.1.1.* *Đặc trưng văn bản* 8](#_Toc484733176)

[*2.1.2.* *Đặc trưng ký hiệu đặc biệt* 9](#_Toc484733177)

[2.2. Kỹ thuật học máy SVM 9](#_Toc484733178)

[***2.2.1.*** ***Ý tưởng*** 9](#_Toc484733179)

[***2.2.2.*** ***Cơ sở lý thuyết*** 10](#_Toc484733180)

[***2.2.3.*** ***Bài toán phân 2 lớp với SVM*** 11](#_Toc484733181)

[***2.2.4.*** ***Các bước chính của phương pháp SVM*** 15](#_Toc484733182)

[*2.2.5.* *Ưu điểm phương pháp SVM trong phân lớp dữ liệu* 15](#_Toc484733183)

[2.3. Kết luận chương 16](#_Toc484733184)

[Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ 17](#_Toc484733185)

[3.1. Các tiêu chuẩn đánh giá 17](#_Toc484733186)

[3.2. Phương pháp thực nghiệm 19](#_Toc484733187)

[***3.2.1.***  ***Biểu diễn văn bản*** 19](#_Toc484733188)

[***3.2.2.*** ***Thư viện hỗ trợ*** 26](#_Toc484733189)

[3.3. Thu thập và mô tả dữ liệu 27](#_Toc484733190)

[*3.3.1.* *Thu thập dữ liệu* 27](#_Toc484733191)

[*3.3.2.* *Mô tả dữ liệu đầu vào* 33](#_Toc484733192)

[3.4. Tiền xử lý dữ liệu 34](#_Toc484733193)

[***3.4.1. Tách từ*** 34](#_Toc484733194)

[***3.4.2. Lọc bộ từ điển*** 35](#_Toc484733195)

[3.5. Kết quả thực nghiệm 36](#_Toc484733196)

[3.7. Kết luận chương 43](#_Toc484733197)

[KẾT LUẬN 44](#_Toc484733198)

[1. Kết quả đạt được 44](#_Toc484733199)

[2. Hạn chế 44](#_Toc484733200)

[3. Hướng phát triển 44](#_Toc484733201)

[DANH MỤC TÀI LIỆU THAM KHẢO 46](#_Toc484733202)

[PHỤ LỤC 47](#_Toc484733203)

DANH MỤC TỪ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Tiếng Anh** | **Tiếng Việt** |
| 1 | SVM | Support vector machine | Máy vector hỗ trợ | |
| 2 | NB | Naïve Bayes | Thuật toán Nave Bayes | |
| 3 | kNN | K–Nearest Neighbor | Thuật toán K – Láng giềng gần nhất | |
| 4 | TF | Term Frequency | Tần số xuất hiện của 1 từ | |
| 5 | IDF | Inverse Document Frequency | Tần số nghịch của 1 từ trong tập văn bản | |
| 6 | Unigram | Unigram | 1-gram | |
| 7 | Bigram | Bigram | 1-gram và 2-gram | |
| 8 | Trigram | Trigram | 1-gram, 2-gram và 3-gram | |
| 9 | API | Application Programming Interface | Giao diện lập trình ứng dụng | |
| 10 | Status | Status | Bài đăng của người dùng trên mạng xã hội Facebook | |
| 11 | Tweet | Tweet | Bài đăng của người dùng trên mạng xã hội Twitter | |
| 12 | NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên | |
| 13 | Facebook | Facebook | Mạng xã hội Facebook | |
| 14 | Twitter | Twitter | Mạng xã hội Twitter | |
| 15 | Youtube | Youtube | Mạng xã hội Youtube | |

DANH MỤC CÁC BẢNG BIỂU

[Hình 1.6: Ví dụ về hồi quy tuyến tính 7](#_Toc484396207)

[Bảng 3.1: Danh sách tập văn bản D gồm 2 câu là C1 và C2 20](#_Toc484396208)

[Bảng 3.2: Danh sách từ điển unigram 21](#_Toc484396209)

[Bảng 3.3: Danh sách từ điển bigram 21](#_Toc484396210)

[Bảng 3.4: Danh sách từ điển trigram 22](#_Toc484396211)

[Bảng 3.5: Danh sách từ điển unigram với trọng số xuất hiện của từ. 24](#_Toc484396212)

[Bảng 3.6: Danh sách từ điển unigram với trọng số TF-IDF. 26](#_Toc484396213)

[Bảng 3.7: Danh sách từ điển unigram với trọng số Binary. 27](#_Toc484396214)

[Bảng 3.8: Thống kế danh sách Status theo người dùng và bài viết 34](#_Toc484396215)

[Bảng 3.9: Thống kế số lượng từ của tập dữ liệu. 36](#_Toc484396216)

[Bảng 3.10: Danh sách các file theo định dạng liblinear. 37](#_Toc484396217)

[Bảng 3.11: Kết quả độ chính xác của tập dữ liệu theo từng Status. 37](#_Toc484396218)

[Bảng 3.12: Kết quả độ chính xác của tập dữ liệu theo từng người dùng. 39](#_Toc484396219)

[Bảng 3.13: Kết quả độ chính xác của tập dữ liệu với 10,000 Status. 42](#_Toc484396220)

[Bảng 3.14: Kết quả độ chính xác của tập dữ liệu với 50,000 Status. 42](#_Toc484396221)

[Bảng 3.15: Kết quả độ chính xác của tập dữ liệu với 100,000 Status. 42](#_Toc484396222)

[Bảng 3.16: Kết quả độ chính xác của tập dữ liệu với 150,000 Status. 43](#_Toc484396223)

DANH MỤC CÁC HÌNH VẼ

[Hình 1.1: Quá trình khớp 7](#_Toc484396377)

[Hình 2.1: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất. 10](#_Toc484396378)

[Hình 2.2: Minh họa bài toán phân 2 lớp bằng phương pháp SVM 12](#_Toc484396379)

[Hình 2.3: Tập dữ liệu được phân chia nhưng có nhiễu 13](#_Toc484396380)

[Hình 2.4: Tập dữ liệu không phân chia tuyến tính 14](#_Toc484396381)

[Hình 2.5: Ví dụ biểu diễn tập dữ liệu trên không gian 2 chiều 15](#_Toc484396382)

[Hình 3.1: Hình ảnh minh họa 10-fold Cross validation. 19](#_Toc484396383)

[Hình 3.2: Graph API cho phép lấy thông tin của người dùng. 29](#_Toc484396384)

[Hình 3.3: Access\_token của người dùng trên Facebook 30](#_Toc484396385)

[Hình 3.4: Minh họa cách lấy danh sách Status trên Facebook. 31](#_Toc484396386)

[Hình 3.5: Tạo project để hỗ trợ lấy nhiều danh sách Status. 32](#_Toc484396387)

[Hình 3.6: Định dạng mỗi dòng trong file csv chứa status lấy được. 32](#_Toc484396388)

[Hình 3.7: File full\_status.csv chứa tất cả Status lấy được. 33](#_Toc484396389)

[Hình 3.8: Minh họa những Status cần phải loại bỏ đi. 34](#_Toc484396390)

[Hình 3.9: Quy trình tách từ. 35](#_Toc484396391)

[Hình 3.10: File vn\_tokenizer\_status.csv chứa danh sách Status sau khi chạy qua vnTokenizer. 36](#_Toc484396392)

[Hình 3.11: Biểu đồ thể hiện kết quả theo trọng số. 38](#_Toc484396393)

[Hình 3.12: Biểu đồ thể hiện kết quả theo tập từ điển. 39](#_Toc484396394)

[Hình 3.13: Biểu đồ thể hiện kết quả theo trọng số của tập dữ liệu theo từng người dùng. 40](#_Toc484396395)

[Hình 3.14: Biểu đồ thể hiện kết quả theo tập từ điển của tập dữ liệu theo từng người dùng. 41](#_Toc484396396)

[Hình 3.15: Biểu đồ kết quả độ chính xác trung bình của từng tập dữ liệu. 44](#_Toc484396397)

MỞ ĐẦU

Trong những năm gần đây, với sự phát triển của các mạng xã hội như: Facebook, Twitter, Youtube…Với số lượng lớn người dùng và liên tục cập nhật thông tin liên quan đến mọi vấn đề như đời sống, xã hội, kinh tế, giải trí… Việc xác định chính xác thông tin cá nhân của người dùng được nhiều tổ chức, công ty, cá nhân quan tâm tới. Trong nhiều trường hợp những thông tin người dùng không cập nhật vào hồ sơ cá nhân hay do người dùng không muốn người khác thấy được vì vậy chúng ta không có đủ thông tin cần thiết. Trong đó, có thông tin quan trọng là giới tính người dùng. Dựa vào một số nghiên cứu đã có, chúng ta có thể dự đoán được giới tính người dùng dựa trên văn phong, cách dùng từ, diễn đạt trong các nội dùng bài viết cùng với việc áp dụng mô hình học máy được huấn luyện trên các bài viết đã biết giới tính của người dùng. Việc dự đoán chính xác giới tính người dùng sẽ đưa ra các số liệu thông kế, các kế hoạch quảng cáo của các công ty, tổ chức cũng như cung cấp các dịch vụ phù hợp với giới tính người dùng trên mạng xã hội nói riêng và mạng internet nói chúng.

Vì vậy, tác giả đã lựa chọn đề tài luận văn thạc sĩ là ***“Dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết”.***

Với mục tiêu đặt ra như vậy, nội dung và kết quả của luận văn được trình bày qua 3 chương như sau:

* Chương 1: Giới thiệu về bài toán dự đoán giới tính và áp dụng để dự đoán giới tính người dùng trên các mạng xã hội Facebook. Phần này cũng đưa ra các phương pháp dự đoán giới tính đã có trong đó chú ý đến phương pháp dựa trên nội dung bài viết là tiền đề để phát triển luận văn.
* Chương 2: Chương này đã giới thiệu chi tiết về phạm vi áp dụng thực nghiệm và đưa ra các đặc trưng sử dụng vào bài toán dự đoán giới tính. Sau đó, chương 2 cũng trình bày chi tiết về kỹ thuật SVM là cơ sở lý thuyết để áp dụng vào thực hiện việc huấn luyện và dự đoán dựa trên nội dung bài viết trên mạng xã hội Facebook.
* Chương 3: Xây dựng các bước để thực nghiệm cho bài toán dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết. Lấy bộ dữ liệu từ các bài viết trên mạng xã hội Facebook, sử dụng thư viện Liblinear có hỗ trợ kỹ thuật học máy SVM. Sau đó đưa bộ dữ liệu vào xử lý và đánh giá kết quả thực nghiệm.

Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH

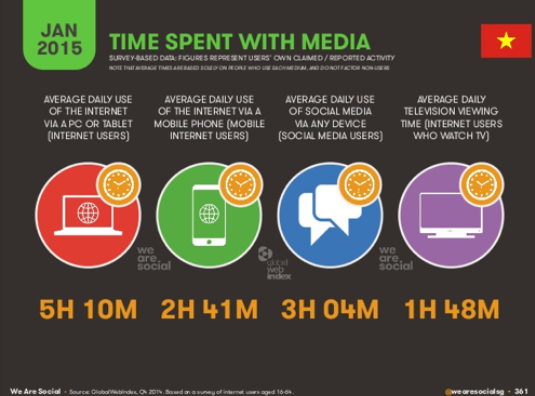
1.1. Giới thiệu bài toán dự đoán giới tính.

1.1.1. Mở đầu

Bài toán dự đoán giới tính là một trong những bài toán kinh điển trong lĩnh vực xử lý dữ liệu văn bản. Bài toán dự đoán giới tính dựa trên nội dung bài viết thực chất là một dạng bài toán phân loại văn bản. Trên thế giới đã có nhiều công trình nghiên cứu đạt những kết quả khả quan về hướng này. Tuy vậy, các nghiên cứu và ứng dụng đối với văn bản tiếng Việt còn có nhiều hạn chế. Phần nhiều lý do là đặc thù của tiếng Việt trên phương diện từ vựng và câu.

Ngày nay, với sự phát triển không ngừng của khoa học công nghệ trên thế giới nói chung và ở Việt Nam nói riêng có những bước tiến vượt bậc. Cơ sở hạ tầng và các trang thiết bị tương đối hiện đại và không ngừng phát triển. Theo báo cáo tổng kết của Bộ TT&TT năm 2016, tỷ lệ người sử dụng Internet ở Việt Nam đạt 62,76% dân số, trong đó tỷ lệ hộ gia đình có truy cập Internet đạt 24,38%, tức là cứ 5 gia đình thì có một hộ sử dụng băng thông rộng cố định. Trong đó, theo thống kê của Cục Viễn thông (Bộ TT&TT) tháng 11/2016, tổng số thuê bao Internet băng rộng cố định đạt hơn 9 triệu thuê bao và số thuê bao băng rộng di động đạt hơn 12,6 triệu thuê bao.

Bên cạnh đó, theo thống kê của “wearesocial.net”, tháng 1-2015, người Việt Nam đang đứng thứ 4 trên thế giới về thời gian sử dụng Internet với 5,2 giờ mỗi ngày, chỉ sau Philippines đứng đầu là 6 giờ, tiếp đó là Thái Lan với 5,5 giờ, và Brazin là 5,4 giờ/ngày.



Hình 1.1 Thời gian sử dụng Internet trung bình một ngày của người Việt Nam

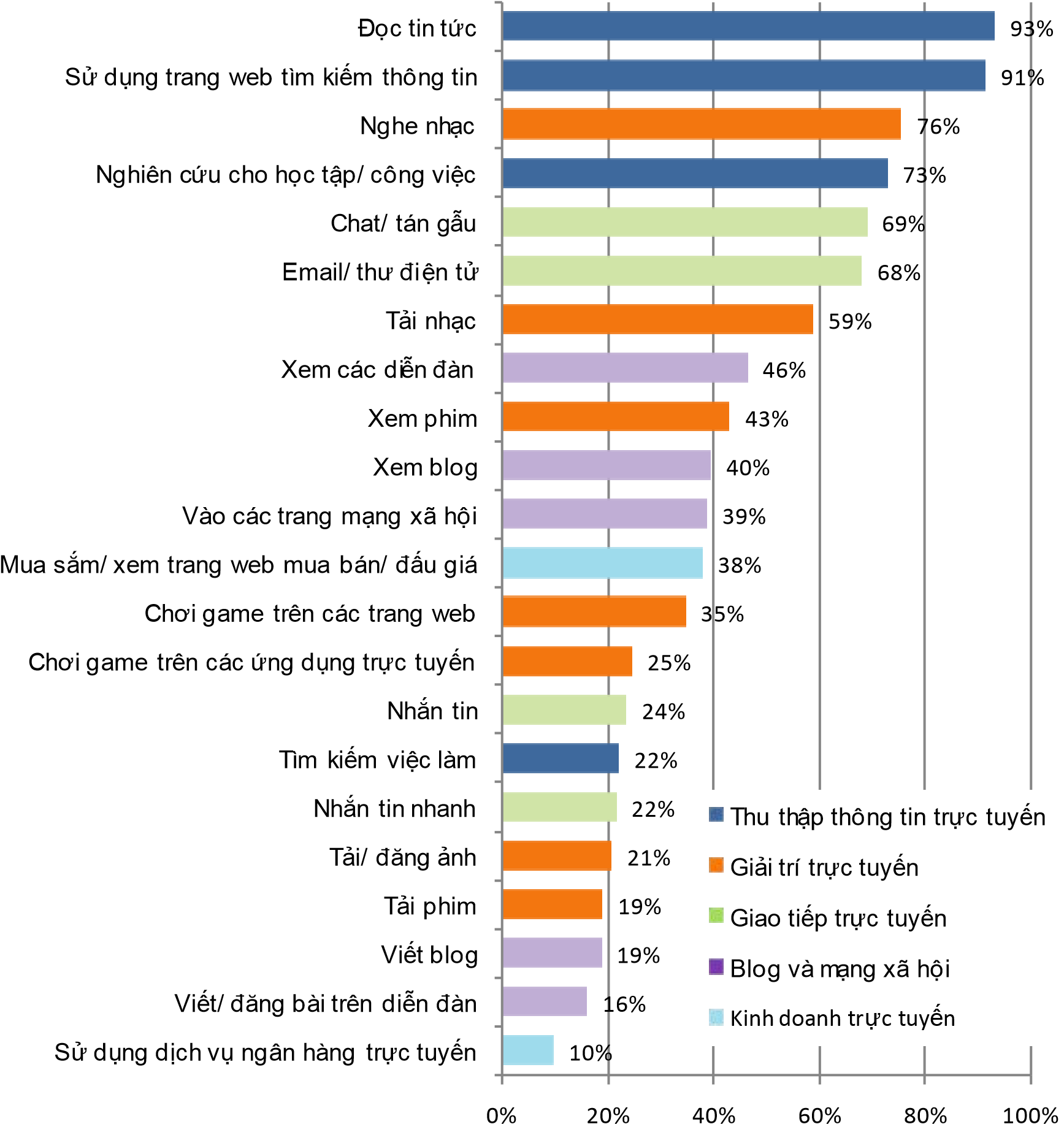
Chính vì sự phát triển không ngừng của công nghệ thông tin và mức độ phổ biến của Internet ngày nay mà thông tin đến với người dùng vô cùng phong phú và liên tục. Người sử dụng Internet hiện nay thường có thói quen truy cập và tìm kiếm đến những các vấn đề mình quan tâm. Hầu hết các thông tin được lưu vào như một phiên làm việc trên mạng. Các thông tin đó có thể là các bài báo, các tài liệu kinh doanh, sản phẩm, các thông tin kinh tế, thương mại điện tử, các thông tin cá nhân khác, ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin để phân loại các thông tin đó cho các mục đích khác nhau như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại...

Với thực tế đó, ta phải xác định và phân loại những thông tin hữu ích từ các nguồn dữ liệu phong phú từ các phiên làm việc, sử dụng Internet của người dùng sao cho phù hợp với đối tượng cụ thể và hỗ trợ các công cụ tự động hoá trợ giúp trong việc phát hiện tri thức và khai thác thông tin.

1.1.2. Bài toán dự đoán giới tính

Nhìn chung, hoạt động thực hiện thường xuyên nhất trên Internet của người dùng là thu thập thông tin, như đọc tin tức hay sử dụng các trang web tìm kiếm. Hơn 90% số lượng người sử dụng Internet đã sử dụng những trang web tìm kiếm, khoảng một nửa trong số họ thậm chí sử dụng hàng ngày. Internet cũng được sử dụng để nghiên cứu cho việc học hay cho công việc bởi một nửa số người sử dụng Internet 1 lần 1 tuần hay thường xuyên hơn. Với các trang web và ứng dụng tương tác trực tuyến mới, người sử dụng không chỉ có cơ hội tìm được thông tin mà cũng đóng góp phần nội dung của riêng họ.

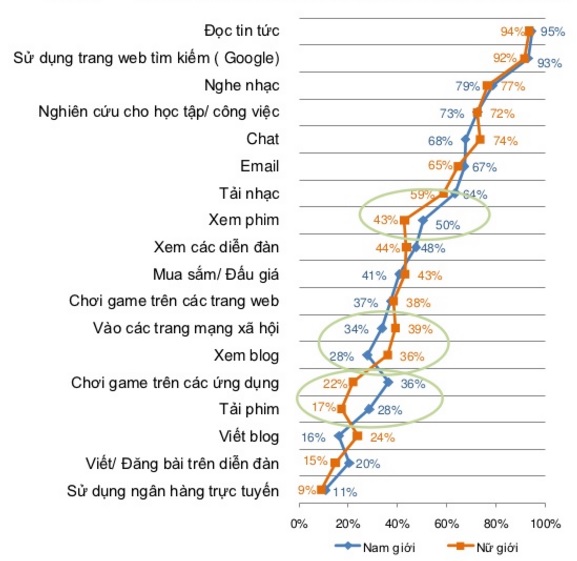
Thương mại điện tử hiện nay, số lượng truy cập đạt mức tăng trưởng đáng kể. Hầu hết các trang phổ biến là các trang web đấu giá và mua bán, nơi có 40% người sử dụng đã từng viếng thăm. Ngân hàng trực tuyến vẫn đang ở giai đoạn sơ khai. Mức độ sử dụng các trang web mua hàng trực tuyến và ngân hàng trực tuyến đã phát triển rất mạnh trong vòng vài năm trở lại đây.



Hình 1.2 Các hoạt động trực tuyến được người dùng mạng sử dụng

(Nguồn: Cimigo NetCitizens)

Việc sử dụng các hoạt động và truy cập Internet có sự khác nhau giữa nam giới và nữ giới. Trung bình một ngày nam giới dành thời gian nhiều hơn cho Internet. Nam giới cũng có một số hoạt động trực tuyến giống với nữ giới. Tuy nhiên có những khác nhau cụ thể ví dự như nam giới có khuynh hướng truy cập những đặc trưng như tin tức thời sự, bóng đá, hay trò chơi và các mặt hàng dành cho nam giới. Trái lại nữ giới thường thích thú với các mục mua sắm, thương mại điện tử, chat và tham gia các trang mạng xã hội và blog.

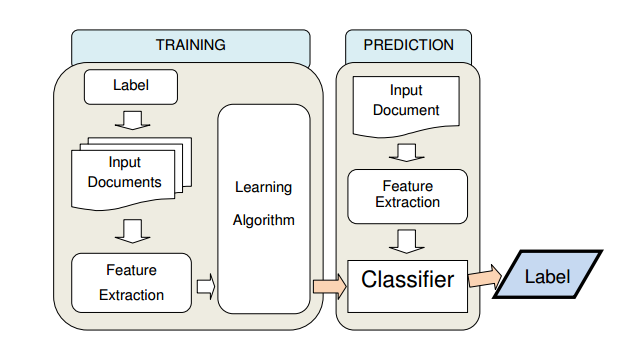


Hình 1.3 Các hoạt động trực tuyến được người dùng mạng theo giới tính

(Nguồn: Cimigo NetCitizens)

Dự đoán giới tính (hay Determination Gender hoặc Gender Prediction) là phương pháp phân loại và xác định các hoạt động được truy cập bởi giới tính Nam hoặc giới tính Nữ từ những hoạt động khác đã biết trước nhãn. Ví dụ một bài báo trong một trang web có thể được truy cập bởi giới tính nam hoặc giới tính nữ (như thể thao, giáo dục, pháp luật, công nghệ thông tin, mỹ phẩm, quần áo …). Việc phân loại có thể được tiến hành một cách thủ công: đọc nội dung của từng hoạt động và gán nó vào một nhãn nào đó. Tuy nhiên, đối với hệ thống gồm rất bản ghi thì phương pháp này sẽ tốn rất nhiều thời gian và công sức. Do vậy cần phải có phương pháp tự động để phân loại giới tính. Phương pháp này giúp cho việc xác định giới tính đạt độ chính xác cao và sử dụng cho các mục đích như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại.

Dưới đây là hình vẽ mô tả quy trình của bài toán xác định giới tính:



Hình 1.4 Quy trình phân loại xác định giới tính

Để tiến hành phân loại xác định giới tính nói chung, chúng ta sẽ thực hiện các bước sau đây:

* Bước 1: Xây dựng bộ dữ liệu huấn luyện dựa trên tập dữ liệu thu thập của người dùng đã được phân loại sẵn. Tiến hành học cho bộ dữ liệu, xử lý và thu thập được dữ liệu của quá trình học là các đặc trưng riêng biệt cho từng nội dung.
* Bước 2: Dữ liệu cần phân loại được xử lý, rút ra các đặc trưng kết hợp với đặc trưng được học trước đó để phân loại và đưa ra kết quả.

Đặc điểm nổi bật của bài toán này là sự đa dạng của hoạt động và đặc trưng của nam giới và nữ giới. Các đặc trưng làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, nếu do con người thực hiện có thể dễ bị nhập nhằng. Ví dụ có hoạt động truy cập về xem thông tin mua sắm quần áo tại một trang web thương mại điện tử, hoạt động truy cập này vẫn có thể được truy cập bời nam giới hoặc nữ giới.

1.1.3. Ứng dụng của bài toán dự đoán giới tính

Trên thế giới đã có một số công trình nghiên cứu với các hướng tiếp cận khác nhau cho bài toán xác định giới tính, bao gồm các tập dữ liệu có đặc trưng thể hiện giới tính. Theo các kết quả trình bày trong các công trình đó thì những cách tiếp cận đều cho kết quả khả quan. Tuy nhiên khó có thể so sánh các kết quả ở trên với nhau vì tập dữ liệu thực nghiệm của mỗi phương pháp là khác nhau.

Hiện nay, công nghệ ngày càng phát triển, đặc biệt với sự ra đời của các trang mạng xã hội, thương mại điện tử nên lượng thông tin lớn, phi cấu trúc, phức tạp, thậm chí là các thông tin rác cũng rất nhiều. Cần thiết phải có những nghiên cứu để xác định được thông tin gì là cần thiết và thông tin nào là dư thừa. Các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin đều đi tìm câu trả lời cho câu hỏi đó. Hầu hết các thông tin đều là các hoạt động trực tuyến như tìm kiếm thông tin, chat, email, mua sắm trực tuyến ... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin của người dùng Internet để phân loại các thông tin đó sao cho phù hợp với giới tính nhằm đưa ra các số liệu thống kê, kế hoạch quảng cáo giúp hệ thống tiếp cận cung cấp thông tin phù hợp, hữu ích cho từng đối tượng.

Trong những năm gần đây, phương pháp phân loại sử dụng Máy vector hỗ trợ (SVM) được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis xây dựng và có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại 2 lớp và đa lớp cũng như trong nhiều ứng dụng khác (như phân loại văn bản theo chủ đề, phát hiện mặt người trong các ảnh, ước lượng hồi quy, dự đoán lỗi phần mềm...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể. Vì những lý do đó mà tôi đã chọn phương pháp này cho việc dự đoán giới tính của người dùng Internet, cụ thể thuật toán và ứng dụng sẽ được trình bày trong các chương sau.

1.2. Các phương pháp dự đoán giới tính

Trên thế giới đã có nhiều phương pháp có thể được sử dụng để dự đoán. Ở giai đoạn đầu phân loại giới tính, hầu hết các nghiên cứu về lĩnh vực này tập trung vào việc nghiên cứu tác giả, đó là những nhiệm vụ xác định hoặc dự đoán các đặc điểm tác giả bằng cách phân tích các câu chuyện, tác phẩm, tiểu thuyết được tạo ra bởi tác giả nam hay tác giả nữ. Các phương pháp mà các nhà nghiên cứu sử dụng trong các nghiên cứu này chủ yếu dựa trên việc phân tích các phong cách viết, văn phong sử dụng các đặc trưng về ngữ pháp chẳng hạn như từ vựng, cú pháp, hoặc các đặc trưng dựa trên nội dung. Như De Vel et al. đã sử dụng 221 đặc trưng để xác định tác giả của email. Argamon và Koppel et al. đã nghiên cứu sự khác biệt trong phong cách viết của nam và nữ trong 604 tài liệu của National Corpus của Anh. Schler et al. khám phá việc sử dụng các đặc trưng và dựa trên nội dung để dự đoán giới tính và độ tuổi của các blogger trên bộ dữ liệu với hơn 71,000 bài viết blog từ blogger.com. Mô hình này đã đạt được kết quả 80% cho dự đoán giới tính và 76% đối với các dự đoán tuổi. Nguyen et al. đã tiến hành một nghiên cứu để dự đoán giới tính và độ tuổi của các thông điệp twitter và diễn đàn bài viết bằng cách sử dụng phương pháp hồi quy với độ chính xác khoảng 80%.

Trong nghiên cứu của Burget et al (2011) đã sử dụng dữ liệu trên mạng xã hội Twitter để huấn luyện và dự đoán giới tính người dùng trên các tweet với từ và ký tự dựa trên đặc trưng n-gram đạt độ chính xác 75.5%. Khi thêm các đặc trưng tên đầy đủ, tên riêng của người dùng, độ chính xác đã tăng lên 89.1%, hơn nữa sử dụng thêm các bài viết mô tả về chính người dùng đã đạt được 92%. Ngoài ra, phương pháp tự huấn luyện khai phá dữ liệu không có nhãn được nghiên cứu nhưng hiệu suất thấp hơn.

Nghiên cứu của Nowson và Oberlander (2006) đạt độ chính xác 92% trong việc dự đoán giới tính chỉ sử dụng đặc trưng n-gram. Dữ liệu của họ gồm 1,400/450 bài đăng viết bởi 47 nữ và 24 nam. Tuy nhiên, đặc trưng n-gram được chọn trước dựa vào việc từ đó có xuất hiện với tần số nhiều trong ngôn ngữ của một giới trong giới khác hay không. Khi tập dữ liệu hoàn chỉnh và được sử dụng để chọn các đặc trưng thì kết quả không đạt được độ chính xác như trước đó.

Yan & Yan (2006) đã sử dụng thuật toán phân loại Naïve Bayes để dự đoán giới tính của các tác giả blog. Trong tổng số 75000 bài viết trên blog cá nhân của 3000 tác giả với giới tính đã được ghi trên trang cá nhân. Kết quả thực nghiệm với độ chính xác 65% và hồi quy 71% đối với giới tính nam.

1.3. Các phương pháp dự đoán giới tính dựa trên các bài biết của người dùng

*1.3.1. Dự đoán giới tính sử dụng bài viết từ blog*

Blog là một loại nhật ký, website cá nhân phổ biến giúp chia sẻ những kinh nghiệm sống hoặc một thông tin gì đó trong cuộc sống hằng ngày của con người. Đây là một loại dữ liệu rất rất lớn chứa các bài viết, văn bản do hàng trăm nghìn tác giả người dùng tạo ra. Những thông tin này chứa đựng rất nhiều các đặc trưng có thể khai thác cho bài toán phân loại, cụ thể ở đây là việc xác định giới tính các blogger. Bài báo nghiên cứu cụ thể về xác định nhân khẩu học và giới tính được Schler et al xây dựng năm 2007 với tập dữ liệu là tất cả blog được truy cập trong một ngày tháng 8 năm 2004.

Nội dung nghiên cứu trú trọng sự khác biệt trong việc viết blog và sự khác biệt giữa nam giới và nữ giới giữa các blogger ở các độ tuổi khác nhau. Các đặc trưng về phong cách và nội dung được đưa ra làm hạt nhân để giải quyết bài toán.

Nghiên cứu sử dụng mô hình MCRW (Multi-Class Real Winnow). Đối với mỗi lớp, , i = 1, ..., m, một vector trọng lượng <, ..., >, trong đó n là kích thước của tập hợp tính năng. Mỗi wi j, được khởi tạo bắt đầu là 1. Các tập huấn luyện được sắp xếp ngẫu nhiên và được xử lý một lần. Thuật toán chạy vòng lặp huấn luyện liên tục, ngẫu nhiên đặt lại các ví dụ sau mỗi chu kỳ. Sau mỗi mười chu kỳ, Thuật toán kiểm tra số lượng các ví dụ đào tạo được phân loại chính xác Nếu con số này đã giảm, thuật toán sẽ quay trở lại. Nếu không có cải tiến nào được tìm thấy sau năm vòng của 10 chu kỳ, thuật toán sẽ được chấm dứt. Nghiên cứu cho thấy mô hình MCRW hiệu quả hơn so với SVM về việc phân loại một số lượng lớn văn bản.

Các kết quả kiểm thử cho thấy được việc phân loại được các blogger theo giới tính theo các nhóm tuổi, kiểu viết và nội dung. Trong các trường hợp được đưa ra, thì sự kết hợp của các đặc trưng phong cách và nội dung cung cấp độ chính xác phân loại tốt nhất.

*1.3.2. Dự đoán giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy*

*a.) Giới thiệu*

Xác định giới tính sử dụng dữ liệu từ các thông điệp Twitter là phương pháp phân loại cho từng bình luận theo đặc trưng dựa trên nội dung bình luận bằng phương pháp hồi quy. Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, ta tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa.

*b). Ý tưởng*

Đọc nội dung twitter của ai đó, trong một số trường hợp, người ta có thể phần nào đoán được giới tính của người dùng. Ví dụ, Bạn có thể biết giới tính người dùng phía sau twitter sau đây?

“Tôi rất thích những bông hoa đặc biệt là hoa hồng và anh ấy cũng vậy <3”

Hồi Quy (regression) là một phương pháp học có giám sát (supervised learning) trong Máy Học. Mục tiêu chính là tìm ra mối quan hệ giữa các đặc trưng của một vấn đề nào đó. Cụ thể hơn, từ một tập dữ liệu cho trước, ta xây dựng một mô hình (phương trình, đồ thị…) khớp nhất với tập dữ liệu, thể hiện được xu hướng biến thiên và mối quan hệ giữa các đặc trưng. Khi có một mẫu dữ liệu mới vào, dựa vào mô hình, chúng ta có thể dự đoán giá trị của mẫu dữ liệu đó. Lấy ví dụ như chúng ta cần dự đoán **giới tính của một Twitter** dựa vào **nội dung** và đặc trưng viết của twitter đó. Như vậy chúng ta cần tìm mối quan hệ giữa **giới tính** phụ thuộc vào **nội dung** và **đặc trưng viết**. Dựa vào tập dữ liệu (giả sử thu thập nội dung, đặc trưng viết và các ký tự đặc biệt của 100 người dùng Twitter), ta xây dựng một phương trình *y* = +​+ trong đó y là giới tính phụ thuộc ​(nội dung) và (đặc trưng viết). Khi có thêm một mẫu dữ liệu của một người dùng mới, chỉ cần áp vào phương trình như vậy ta sẽ dự đoán được giới tính của người đó.

Ta thấy phương trình *y* = +​+​ là phương trình của mặt phẳng trong không gian 3 chiều. Những mô hình tương tự như phương trình đường thẳng, phương trình mặt phẳng chính là những mô hình tuyến tính. Hồi quy tuyến tính (linear regression) là một mô hình đơn giản trong bài toán hồi quy, trong đó chúng ta dùng đường thẳng, mặt phẳng, hay phương trình tuyến tính nói chung để dự đoán xu hướng của dữ liệu. Giải bài toán hồi quy tuyến tính chính là đi tìm các tham số , ...để xác định phương trình tuyến tính.



Hình 1.6: Ví dụ về hồi quy tuyến tính

1.4. Kết luận chương

Chương 1 đã giới thiệu về bài toán dự đoán giới tính các phương pháp có thể dự đoán giới tính người dùng và trình bày một số bài báo đã có về dự đoán giới tính dựa trên các nội dung bài viết khác nhau. Đây là tiền đề tham khảo để phát triển luân văn.

Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI

2.1. Các đặc trưng sử dụng

Trong việc phân loại giới tính người dùng mạng ta sử dụng chủ yếu là Status. Status sẽ có các đặc trưng như đặc trưng về text và đặc trưng về ký tự.

*2.1.1. Đặc trưng văn bản*

Tiếng Việt là ngôn ngữ đơn lập. Đặc điểm này bao quát tiếng Việt cả về mặt ngữ âm, ngữ nghĩa, ngữ pháp. Khác với các ngôn ngữ châu Âu, mỗi từ là một nhóm các ký tự có nghĩa được cách nhau bởi một khoảng trắng. Còn tiếng Việt, và các ngôn ngữ đơn lập khác, thì khoảng trắng không phải là căn cứ để nhận diện từ.

Tiếng:

* Trong tiếng Việt trước hết cần chú ý đến đơn vị xưa nay vẫn quan gọi là tiếng. Về mặt ngữ nghĩa, ngữ âm, ngữ pháp, đều có giá trị quan trọng.
* Sử dụng tiếng để tạo từ có hai trường hợp:
* Trường hợp một tiếng: đây là trường hợp một tiếng được dùng làm một từ, gọi là từ đơn. Tuy nhiên không phải tiếng nào cũng tạo thành một từ.
* Trường hợp hai tiếng trở lên: đây là trường hợp hai hay nhiều tiếng kết hợp với nhau, cả khối kết hợp với nhau gắn bó tương đối chặt chẽ, mới có tư cách ngữ pháp là một từ. Đây là trường hợp từ ghép hay từ phức.

Từ:

* Có rất nhiều quan niệm về từ trong tiếng Việt, từ nhiều quan niệm về từ tiếng Việt khác nhau đó chúng ta có thể thấy đặc trưng cơ bản của "từ" là sự hoàn chỉnh về mặt nội dung, từ là đơn vị nhỏ nhất để đặt câu.
* Người ta dùng "từ" kết hợp thành câu chứ không phải dùng "tiếng", do đó quá trình tách câu thành các "từ" cho kết quả tốt hơn là tách câu bằng “tiếng”.

*2.1.2. Đặc trưng ký hiệu đặc biệt*

Người dùng mạng xã hội không chỉ dùng những câu từ để thể hiện cảm xúc, suy nghĩ, bày tỏ thái độ về một vấn đề gì đó họ còn có thể dùng những ký hiệu đặc biệt những biểu tưởng cảm xúc để làm việc đó. Mỗi mạng xã hội đều có sẵn một tập hợp các biểu tượng cho phép người dùng sử dụng trong các bài viết đây là những dấu hiệu quan trọng trong việc giao tiếp không chính thức. Biểu tượng cảm xúc là những dấu hiệu quan trọng trong việc giao tiếp không chính thức. Các nghiên cứu cho thấy rằng nữ giới sử dụng cao gấp 3,5 lần so với nam giới. Điều này phù hợp với các phát hiện trong nghiên cứu tiếng nói, trong đó tiếng cười trong các cuộc nói chuyện thường là nữ. Nam giới và phụ nữ cũng khác nhau trong các loại biểu tượng cảm xúc được sử dụng. Trong khi cả đàn ông và phụ nữ thích nụ cười (:)), phụ nữ có thể sử dụng các biểu tượng cảm xúc như trái tim (<3) và nam giới có khuynh hướng nở nụ cười (: D) và nháy mắt (;)).

2.2. Kỹ thuật học máy SVM

Kỹ thuật học máy SVM là viết tắt của từ Support Vector Machine, đây một phương pháp trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy. SVM dạng chuẩn nhận dữ liệu vào và phân loại chúng vào 2 lớp khác nhau.

***2.2.1. Ý tưởng***

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector, trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp + và lớp -. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất, điều này được minh họa như sau:



Hình 2.1: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất.

***2.2.2. Cơ sở lý thuyết***

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

Cho tập mẫu (x1, y1), (x2, y2),…, (xf, yf)} với xi ∈ Rn, thuộc vào hai lớp nhãn: yi ∈ {-1,1} là nhãn lớp tương ứng của các xi (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vectơ trong không gian:

Như vậy, f(Xi) biểu diễn sự phân lớp của Xi vào hai lớp như đã nêu. Ta nói yi = +1 nếu Xi thuộc lớp I và yi = -1 nếu Xi thuộc lớp II. Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau: Tìm min w với W thỏa mãn điều kiện sau:

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi về thành dạng đẳng thức. Một đặc điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc các Support Vector và nó có khoảng cách đến mặt phẳng quyết định là . Cho dù các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

***2.2.3. Bài toán phân 2 lớp với SVM***

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới thì cần phải xác định được phân vào lớp +1 hay lớp -1.

Ta xét 3 trường hợp, mỗi trường hợp sẽ có 1 bài toán tối ưu, giải được bài toán tối ưu đó ta sẽ tìm được siêu phẳng cần tìm.

**Trường hợp 1:**

Tập D có thể phân chia tuyến tính được mà không có nhiễu (tất cả các điểm được gán nhãn +1 thuộc về phía dương của siêu phảng, tất cả các điểm được gán nhãn -1 thuộc về phía âm của siêu phẳng).



Hình 2.2: Minh họa bài toán phân 2 lớp bằng phương pháp SVM

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được.

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

Ta sẽ tìm siêu phẳng tách với 𝑤 ∈ ℝ𝑛 là vector trọng số, 𝑏 ∈ ℝ𝑛 là hệ số tự do, sao cho:



Lúc này ta cần giải bài toán tối ưu:



**Trường hợp 2:**

Tập dữ liệu D có thể phân chia tuyến tính được nhưng có nhiễu. Trong trường hợp này, hầu hết các điểm đều được phân chia đúng bởi siêu phẳng. Tuy nhiên có 1 số điểm bị nhiễu, nghĩa là: Điểm có nhãn dương nhưng lại thuộc phía âm của siêu phẳng, điểm có nhãn âm nhưng lại thuộc phía dương của siêu phẳng.



Hình 2.3: Tập dữ liệu được phân chia nhưng có nhiễu

Trong trường hợp này, ta sử dụng 1 biến mềm  sao cho: 



Bài toán tối ưu trở thành:



Trong đó C là tham số xác định trước, định nghĩa giá trị ràng buộc, C càng lớn thì mức độ vi phạm đối với những lỗi thực nghiệm (là lỗi xảy ra lúc huấn luyện, tính bằng thương số của số phần tử lỗi và tổng số phần tử huấn luyện) càng cao.

**Trường hợp 3:**

Tập dữ liệu D không thể phân chia tuyến tính được, ta sẽ ánh xạ các vector dữ liệu x từ không gian n chiều vào một không gian m chiều (m>n), sao cho trong không gian m chiều, D có thể phân chia tuyến tính được.



Hình 2.4: Tập dữ liệu không phân chia tuyến tính

Gọi  là một ánh xạ phi tuyến từ không gian  vào không gian .



Bài toán tối ưu trở thành:



Ví dụ:

Để dễ hiểu hơn ta xét ví dụ mô tả hình học sau: Xét trong không gian 2 chiều (n=2), tập dữ liệu được cho bởi tập các điểm trên mặt phẳng.



Hình 2.5: Ví dụ biểu diễn tập dữ liệu trên không gian 2 chiều

Bây giờ ta tiến hành tìm siêu phẳng phân lớp dựa trên phương pháp SVM (1). Ta sẽ tìm 2 siêu phẳng song song (nét đứt trong hình) sao cho khoảng cách giữa chúng là lớn nhất để có thể phân tách lớp này thành 2 phía (Ta gọi là 2 siêu phẳng phân tách). Siêu phẳng (1) nằm giữa 2 siêu phẳng trên (nét đậm trong hình).

Hình trên cho ta tập dữ liệu có thể phân tách tuyến tính. Bây giờ ta xét trường hợp tập dữ liệu không thể phân tách tuyến tính. Bây giờ ta sẽ xử lý bằng cách ánh xạ tập dữ liệu đã cho vào một không gian mới có số chiều lớn hơn không gian cũ (Gọi là không gian đặc trưng) mà trong không gian này tập dữ liệu có thể phân tách tuyến tính. Trong không gian đặc trưng ta sẽ tiếp tục tìm 2 siêu phẳng phân tách như trường hợp ban đầu.

Các điểm nằm trên 2 siêu phẳng phân tách gọi là các vector hỗ trợ (Support vector). Các điểm này quyết định hàm phân tách dữ liệu. Từ đây, chúng ta có thể thấy phương pháp SVM không phụ thuộc vào các mẫu dữ liệu ban đầu, mà chỉ phụ thuộc vào các suport vector (quyết định 2 siêu phẳng phân tách). Cho dù các điểm khác bị xoá thì thuật toán vẫn cho ra các kết quả tương tự. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác do các điểm trong tập dữ liệu đều được dùng để tối ưu kết quả.

***2.2.4. Các bước chính của phương pháp SVM***

Phương pháp SVM yêu cầu dữ liệu được biểu diễn như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM.

Tiền xử lý dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên chuẩn hóa dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].

Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.

Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng đụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.

Sử dụng các tham số cho việc huấn luyện với tập mẫu. Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

*2.2.5. Ưu điểm phương pháp SVM trong phân lớp dữ liệu*

Như đã biết, phân lớp dữ liệu là một tiến trình đưa các dữ liệu chưa biết nhãn vào các lớp dữ liệu đã biết nhãn tương ứng. Mỗi nhãn được xác định bởi một số tập dữ liệu mẫu của nhãn đó. Để thực hiện quá trình phân lớp, các phương pháp huấn luyện được sử dụng để xây dựng tập phân lớp từ các bản ghi mẫu, sau đó dùng tập phân lớp này để dự đoán lớp của những bản ghi mới chưa biết nhãn.

Chúng ta có thể thấy các thuật toán phân lớp hai lớp như SVM đều có đặc điểm chung là yêu cầu dữ liệu phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các uớc lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp thì SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất (hơn 10.000 chiều) trong khi đó các phương pháp khác có số chiều bé hơn nhiều (như Naïve Bayes là 2000, k-Nearest Neighbors là 2415…).

Trong công trình của mình năm 1999, Joachims đã so sánh SVM với Naïve Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003, Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính được đề cập trước đây của tập dữ liệu. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác.

Theo Xiaojin Zhu thì trong các công trình nghiên cứu của nhiều tác giả (chẳng hạn như Kiritchenko và Matwin vào năm 2001, Hwanjo Yu và Han vào năm 2003, Lewis vào năm 2004) đã chỉ ra rằng thuật toán SVM đem lại kết quả tốt nhất phân lớp văn bản.

2.3. Kết luận chương

Chương 2 của luận văn tập trung vào trình bày kỹ thuật học máy SVM cơ sở lý thuyết và áp dụng trong bài toán dự đoán giới tính chính là bài toán phân 2 lớp của SVM là tiền đề để đánh giá với dữ liệu thực nghiệm.

Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Thu thập và mô tả dữ liệu

*3.1.1. Thu thập dữ liệu*

Dữ liệu sử dụng trong luận văn sẽ lấy các nội dùng bài viết của người dùng trên mạng xã hội Facebook. Nội dung bài viết lấy để làm dữ liệu thực nghiệm trong luận văn là danh sách Status (trạng thái) cho phép người dụng thông báo cho bạn bè họ đang ở đâu làm gì, trong Status thì có thể chèn thêm hình ảnh, đường link hoặc video, tag bạn bè…

Trên Facebook có cung cấp **Graph API** cho phép lấy những thông tin người dùng trong đó có các bài Status của họ và bạn bè.



Hình 3.2: Graph API cho phép lấy thông tin của người dùng.

Để sử dụng API này người dùng cần phải có access gọi là **access\_token**

**Access\_token** là của app trên facebook hoặc của tài khoản facebook.



Hình 3.3: Access\_token của người dùng trên Facebook

Khi có access\_token em có thể lấy được các bài Status của bạn bè, tuy nhiên không phải ai là bạn bè cũng có thể lấy được các thông tin cần thiết. Facebook có cơ chế chỉ cho phép truy cập thông tin người dùng đã cho phép **Graph API** truy cập.

Ví dụ: Facebook của em có tầm 1.600 bạn bè, thì chỉ lấy được bài Status của 150 bạn bè.



Hình 3.4: Minh họa cách lấy danh sách Status trên Facebook.

Để lấy được nhiều bài Status em có tạo 1 project java là **SocialNetworkData**

Trong project này em có sử dụng thư viện hỗ trợ là **restfb-1.33.0.jar** là một open source cho phép gọi các APIs của **Graph API** để lấy thông tin.

Input: Là **access\_token** của tài khoản người dùng Facebook.

Output: Là file csv có chứa danh sách Status (xóa các dấu cách thừa và dấu ‘,’ và xuống dòng) với mỗi dòng là một Status

Để có sự đánh giá độ chính xác của phương pháp SVM em chỉ lấy dữ dữ liệu người dùng đã có thông tin về giới tính rõ ràng (nam/nữ), chỉ lấy Status là văn bản thuần không chứa URL, tag bạn bè, hình ảnh, video…



Hình 3.5: Tạo project để hỗ trợ lấy nhiều danh sách Status.

Mỗi dòng trong file csv sẽ có định dạng như sau:

< Id người dùng>, <Tên người dùng>, <Giới tính người dùng>, < Id Status>, < Status>



Hình 3.6: Định dạng mỗi dòng trong file csv chứa status lấy được.

Số lượng Status lấy được lưu vào file **full\_status.csv.**



Hình 3.7: File full\_status.csv chứa tất cả Status lấy được.

File full\_status.csv hiện tại có nhiều Status cần loại bỏ như sau:

* Có số lượng từ ký tự (ngăn cách nhau bằng dấu cách) nhỏ hơn 5 hoặc lớn hơn 225.
* Các Status trùng nhau. Có nhiều người cùng đăng Status sưu tầm, câu truyện…
* Các Status không phải tiếng Việt. Facebook là mạng xã hội nhiều người dùng ở các quốc gia, vùng miền khác nhau. Facebook ở Việt Nam thường có lẫn những Status tiếng Trung, Tiếng Nhật, tiếng Nga… Với những Status là tiếng Việt không dấu cũng cần phải loại bỏ.
* Những Status có quá nhiều kí tự hơn từ.





Hình 3.8: Minh họa những Status cần phải loại bỏ đi.

Sau đó loại bỏ những Staus không phù hợp em sẽ lưu danh sách Status còn lại vào file có tên là **full\_status\_filter.csv**.

*3.1.2. Mô tả dữ liệu đầu vào*

Trong file **full\_status\_filter.csv** có chứa danh sách Status của nhiều người dùng khác nhau.

Bảng 3.8 là thống kế tập dữ liệu đầu vào theo người dùng và theo Status:

* Với thống kế theo từng người dùng ta coi một người dùng có nhiều Status, tập hợp các Status thế hiện giới tính của người dùng đó.
* Với thống kế theo từng Status thì mỗi Status thể hiện một giới tính của người dùng, các Status của cùng người dùng là riêng biệt nhau khi đánh giá theo bài viết.

Bảng 3.8: Thống kế danh sách Status theo người dùng và bài viết

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Người dùng** | | **Status** | |
| **Số lượng** | **Tỉ lệ** | **Số lượng** | **Tỉ lệ** |
| **Nam** | 659 | 57.8% | 109,170 | 49.7% |
| **Nữ** | 482 | 42.2% | 107,702 | 50.3% |
| **Tổng số** | 1,141 | 100% | 216,872 | 100% |

Nhận xét: Từ bảng 3.8 thấy rằng nữ giới viết nhiều Status hơn nam giới. Trung bình một người dùng viết tầm 190 trong đó một nam giới là 163 Status và nữ giới có 226 Status.

3.2. Các tiêu chuẩn đánh giá

Việc đánh giá một giải thuật học máy cho bộ dữ liệu là rất quan trọng, nó cho phép đánh giá được độ chính xác của các kết quả phân lớp. Đánh giá còn giúp so sánh các giải thuật học máy khác nhau. Việc đánh giá độ chính xác của một giải thuật học máy thường được thực hiện dựa trên thực nghiệm hơn là dựa trên phân tích.

Đánh giá độ chính xác thường phụ thuộc vào các yếu tố sau:

* Tập dữ liệu càng lớn thì độ chính xác càng tốt.
* Tập kiểm thử cần lớn thì việc đánh giá càng chính xác.
* Vấn đề là rất khó (ít khi) có thể có được các tập dữ liệu (rất) lớn.

Để đánh giá một giải thuật máy học một số chỉ số thông dụng được sử dụng. Giả sử như bộ phân lớp có 2 lớp là lớp âm (negative) và lớp dương (positive) thì các chỉ số được định nghĩa như sau:

* + TP- True positive: số phần tử dương được phân loại dương.
  + FN - False negative: số phần tử dương được phân loại âm.
  + TN- True negative: số phần tử âm được phân loại âm.
  + FP - False positive: số phần tử âm được phân loại dương.
  + Độ chính xác (Accuracy) = .

Chỉ số Accuracy sẽ được sử dụng trong luận văn để quá trình đánh giá kết quả thực nghiệm.

Có nhiều phương pháp để đánh giá một giải thuật học máy trong đó có phương pháp phổ biến hay được sử dụng là Cross-validation.

Cross-validation là một phương pháp kiểm tra độ chính xác của một giải thuật máy học dựa trên tập dữ liệu cho trước. Thay vì chỉ dùng một phần dữ liệu làm tập huấn luyện và kiểm tra thì phương pháp Cross-validation dùng toàn bộ tập dữ liệu để đánh giá.

Có 3 phương pháp Cross-validation phổ biến là:

* **Phương pháp Hold-out**: là phương pháp đơn giản nhất. Dữ liệu được chia một cách ngẫu nhiên thành một tập dữ liệu huấn luyện và một tập dữ liệu kiểm tra. Dùng tập đầu tiên để huấn luyện rồi dùng ngay tập còn lại để kiểm tra.
* **Phương pháp K-fold**: đây là phương pháp nâng cấp của hold-out. Toàn bộ dữ liệu được chia thành **k** tập con không giao nhau. Quá trình học của máy có **k** lần. Trong mỗi lần, một tập con được dùng để kiểm thử và (**k-1**) tập còn lại dùng để huấn luyện. Các lựa chọn thông thường của **k** là 5 hoặc 10. Độ chính xác cuối cùng bằng trung bình độ chính xác của k lần học.



Hình 3.1: Hình ảnh minh họa 10-fold Cross validation.

* Phương pháp **Leave-one-out**: Tương tự như k-fold nhưng tối đa hóa số tập con (k = số dữ liệu).

Trong luận văn này sẽ sử dụng phương pháp **k-fold Cross validation** với **10-fold** để thực hiện việc đánh giá.

3.3. Phương pháp thực nghiệm

***3.3.1. Biểu diễn văn bản***

Chúng ta cần biểu diễn văn bản một vector của các feature để dùng được giải thuật SVM để phân loại. Trước tiên cần xây dựng bộ từ điển cho tập dữ liệu văn bản. Trong luận văn này sẽ sử dụng mô hình n-gram để xây dựng bộ từ điển.

Gram ở đây là đơn vị nhỏ nhất – hay nói cách khác trong câu thì nó chỉ bao gồm một từ. Một cụm n-gram là một dãy con gồm n-yếu tố liên tiếp nhau của một dãy các yếu tố cho trước. Yếu tố ở đây có thể là âm tiết, chữ cái hoặc từ vựng... Nhãn từ loại và các n-gram thường được thu thập từ một văn bản hoặc lời nói. Số phần tử trong một n-gram được gọi là bậc của n-gram, thông thường n-gram có bậc từ 1 tới 3:

* 1-gram là n-gram bậc 1 hay được gọi là unigram
* 2-gram là n-gram bậc 2 còn được gọi là bigram
* 3-gram là n-gram bậc 3 hay được gọi là trigram

N-gram được dùng để ước lượng xác suất xuất hiện của một yếu tố dựa vào các yếu tố xung quanh nó trong câu. Do đó, n-gram có thể áp dụng cho các hệ thống tách từ, gán nhãn từ loại, phát hiện lỗi chú giải từ loại…

Ví dụ cho tập văn bản D gồm 2 câu C1 và C2 như bảng sau:

Bảng 3.1: Danh sách tập văn bản D gồm 2 câu là C1 và C2

|  |  |  |  |
| --- | --- | --- | --- |
| **Số thứ tự** | **Giới tính** | **Mã câu** | **Nội dung** |
| 1 | Nữ | C1 | Con mèo ngồi trên chiếc mũ |
| 2 | Nam | C2 | Con chó cắn con mèo và chiếc mũ |

Tập từ điển tương ứng với n-gram như sau:

* 1-gram: con, mèo, ngồi, trên, chiếc, mũ, chó, cắn, và.
* 2-gram: con mèo, mèo ngồi, ngồi trên, trên chiếc, chiếc mũ, con chó, chó cắn, cắn con, mèo và, và chiếc.
* 3-gram: con mèo ngồi, mèo ngồi trên, ngồi trên chiếc, trên chiếc mũ, con chó cắn, chó cắn con, cắn con mèo, con mèo và, mèo và chiếc, và chiếc mũ.

Để có thể sử dụng được các thuật toán học máy cho văn bản, việc xây dựng tập từ điền để biểu diễn văn bản là rất quan trọng nó ảnh hướng đến kết quả dự đoán, phân loại. Dựa vào mô hình n-gram em sẽ xây dựng tập danh sách từ điển đối với tập dữ liệu đầu thành 3 tập từ điển để đánh giá.

**Tập từ điển unigram:** Là tập hợp danh sách từ điển chỉ có 1-gram

Ví dụ tập văn bản D ở Bảng 3.1 gồm danh sách 9 từ như sau:

Bảng 3.2: Danh sách từ điển unigram

|  |  |
| --- | --- |
| **Thứ tự** | **Từ** |
| 1 | con |
| 2 | mèo |
| 3 | ngồi |
| 4 | trên |
| 5 | chiếc |
| 6 | mũ |
| 7 | chó |
| 8 | cắn |
| 9 | và |

**Tập từ điển bigram:** Là tập hợp danh sách từ gồm 1-gram và 2-gram.

Ví dụ tập văn bản D ở Bảng 3.1 gồm danh sách 19 từ như sau:

Bảng 3.3: Danh sách từ điển bigram

|  |  |
| --- | --- |
| **Thứ tự** | **Từ** |
| 1 | con |
| 2 | mèo |
| 3 | ngồi |
| 4 | trên |
| 5 | chiếc |
| 6 | mũ |
| 7 | chó |
| 8 | cắn |
| 9 | và |
| 10 | con mèo |
| 11 | mèo ngồi |
| 12 | ngồi trên |
| 13 | trên chiếc |
| 14 | chiếc mũ |
| 15 | con chó |
| 16 | chó cắn |
| 17 | cắn con |
| 18 | mèo và |
| 19 | và chiếc |

**Tập từ điển trigram:** Là tập hợp dang sách từ gồm 1-gram, 2-gram và 3-gram.

Ví dụ tập văn bản D ở Bảng 3.1 gồm danh sách 29 từ như sau:

Bảng 3.4: Danh sách từ điển trigram

| **Thứ tự** | **Từ** |
| --- | --- |
| 1 | con |
| 2 | mèo |
| 3 | ngồi |
| 4 | trên |
| 5 | chiếc |
| 6 | mũ |
| 7 | chó |
| 8 | cắn |
| 9 | và |
| 10 | con mèo |
| 11 | mèo ngồi |
| 12 | ngồi trên |
| 13 | trên chiếc |
| 14 | chiếc mũ |
| 15 | con chó |
| 16 | chó cắn |
| 17 | cắn con |
| 18 | mèo và |
| 19 | và chiếc |
| 20 | con mèo ngồi |
| 21 | mèo ngồi trên |
| 22 | ngồi trên chiếc |
| 23 | trên chiếc mũ |
| 24 | con chó cắn |
| 25 | chó cắn con |
| 26 | cắn con mèo |
| 27 | con mèo và |
| 28 | mèo và chiếc |
| 29 | và chiếc mũ |

Sau khi đã xây dựng được tập từ điển, để biểu diễn văn bản chúng ta cần tìm trọng số cho tập từ điển. Trong luận văn sẽ sử dụng 3 trọng số là: số lần xuất hiện của từ, chỉ số TF-IDF, và trọng số Binary

**Bài toán**

* Input: Cho một tập văn bản gồm m văn bản D = {, , ..., } và T là một tập từ điển gồm n từ khác nhau T = {, , ..., }.
* Output: Xây dựng w = () là ma trận trọng số, trong đó là trọng số của từ ∈ T trong văn bản ∈ D.

***a). Trọng số xuất hiện của từ (count)***

Trọng số này được xác định bằng cách đếm số lần xuất hiện của từ ∈ T trong văn bản ∈ D.

.

Trong ví dụ ở Bảng 3.1 sử dụng tập từ điển Bảng 3.2. Trong câu C1, “con”, “mèo”, “ngồi”, “trên”, “chiếc” và “mũ” mỗi từ xuất hiện 1 lần. Trong câu C2, “con” xuất hiện 2 lần và “mèo”, “chiếc”, “mũ”, “chó”, “cắn” và “và” mỗi từ xuất hiện 1 lần. Như vậy trọng số cho C1 và C2 sẽ là:

C1: {1, 1, 1, 1, 1, 1, 0, 0, 0}

C2: {2, 1, 0, 0, 1, 1, 1, 1, 1}

Bảng 3.5: Danh sách từ điển unigram với trọng số xuất hiện của từ.

|  |  |  |  |
| --- | --- | --- | --- |
| **Thứ tự** | **Từ** | **C1** | **C2** |
| 1 | con | 1 | 2 |
| 2 | mèo | 1 | 1 |
| 3 | ngồi | 1 | 0 |
| 4 | trên | 1 | 0 |
| 5 | chiếc | 1 | 1 |
| 6 | mũ | 1 | 1 |
| 7 | chó | 0 | 1 |
| 8 | cắn | 0 | 1 |
| 9 | và | 0 | 1 |

***b). Trọng số TF-IDF***

TF-IDF viết tắt của Term Frequency – Inverse Document Frequency, là trọng số của một từ thu được qua [thống kê](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA) thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

Trọng số TF-IDF được tính như sau:

TF-IDF(, , D) = TF(, ) x IDF(, D).

Trong đó:

* TF: dùng để ước lượng tần xuất xuất hiện của từ trong văn bản. Tuy nhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn. Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó). Được tính như công thức sau:

TF(, ) =

* IDF: dùng để ước lượng mức độ quan trọng của từ đó như thế nào. Khi tính tần số xuất hiện TF thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn. Ví dụ: từ nối (và, nhưng…), giới từ (ở, trong, trên..), từ chỉ định (ấy, đó, nhỉ..)... Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF. Được tính như công thức sau:

IDF(, D) = log()

Trong ví dụ ở Bảng 3.1 sử dụng tập từ điển Bảng 3.2. Trọng số TF-IDF cho từ “con” trong C1 được tính như sau:

* TF(“con”, C1) = = 0.1667
* IDF (“con”, D) = log() = 0
* TF-IDF(“con”, C1, D) = TF(“con”, C1) x IDF(“con”, D) = 0.1667 x 0 = 0

Trọng số TF-IDF cho từ “chó” trong C2 được tính như sau:

* TF(“chó”, C2) = = 0.125
* IDF (“chó”, D) = log() = 0.301
* TF-IDF(“chó”, C2, D) = TF(“chó”, C2) x IDF(“chó”, D) = 0.125 x 0.301 = 0.038

C1: {0, 0, 0.05, 0.05, 0, 0, 0, 0}

C2: {0, 0, 0, 0, 0, 0.038, 0.038, 0.038}

Bảng 3.6: Danh sách từ điển unigram với trọng số TF-IDF.

|  |  |  |  |
| --- | --- | --- | --- |
| **Thứ tự** | **Từ** | **C1** | **C2** |
| 1 | con | 0 | 0 |
| 2 | mèo | 0 | 0 |
| 3 | ngồi | 0.05 | 0 |
| 4 | trên | 0.05 | 0 |
| 5 | chiếc | 0 | 0 |
| 6 | mũ | 0 | 0 |
| 7 | chó | 0 | 0.038 |
| 8 | cắn | 0 | 0.038 |
| 9 | và | 0 | 0.038 |

***c). Trọng số Binary***

Trọng số binary quan tâm đến sự xuất hiện hay không xuất hiện của từ trong câu. Nếu xuất hiện giá trị là 1 ngược lại nếu không xuất hiện trọng số là 0.

Trong ví dụ ở Bảng 3.1 sử dụng tập từ điển Bảng 3.2. Trong C1 có các từ “con”, “mèo”, “ngồi”, “trên”, “chiếc” và “mũ”. Trong C2 có các từ “con”, “mèo”, “chiếc”, “mũ”, “chó”, “cắn” và “và”. Như vậy trọng số cho C1 và C2 sẽ là:

C1: {1, 1, 1, 1, 1, 1, 0, 0, 0}

C2: {1, 1, 0, 0, 1, 1, 1, 1, 1}

Bảng 3.7: Danh sách từ điển unigram với trọng số Binary.

|  |  |  |  |
| --- | --- | --- | --- |
| **Thứ tự** | **Từ** | **C1** | **C2** |
| 1 | con | 1 | 1 |
| 2 | mèo | 1 | 1 |
| 3 | ngồi | 1 | 0 |
| 4 | trên | 1 | 0 |
| 5 | chiếc | 1 | 1 |
| 6 | mũ | 1 | 1 |
| 7 | chó | 0 | 1 |
| 8 | cắn | 0 | 1 |
| 9 | và | 0 | 1 |

***3.3.2. Thư viện hỗ trợ***

Có nhiều thư viện hỗ trợ phương pháp học máy SVM trong đó có bộ thư viện Liblinear. Thư viện này hỗ trợ phương pháp học máy SVM và có ưu điểm nổi bật như sau:

* Tốc độ xử lý rất nhanh.
* Có thể phân loại những bài toán có từ hàng triệu đến hàng chục triệu đặc trưng.
* Yêu cầu cấu hình máy thấp, máy tính cá nhân thông thường cũng có thể hoạt động được.

- *Định dạng file*: Định dạng của file dữ liệu huấn luyện và file kiểm tra là:

<label><index1>:<value1><index2>:<value2> ...

Trong đó:

<label>: là giá trị đích của tập huấn luyện. Đối với việc phân lớp là một số nguyên xác định một lớp. Với bài toán dự đoán giới tính thì label sẽ có hai giá trị là 1 nếu là nam và là -1 nếu là nữ

<index>: là một số nguyên bắt đầu từ 1. Là thứ tự từ trong bộ từ điển.

<value>: là trọng số của index. Nếu value = 0 thì không cần phải ghi.

- *Cách sử dụng*: Trong luận văn em sử dụng kỹ thuật đánh giá 10-fold Cross validation thì chỉ cần dùng train với câu lệnh như sau:

train –v 10 training\_set\_file

Trong đó:

* training\_set\_file: là file huấn luyện
* -v 10: có nghĩa là sử dụng 10-fold Cross validation

Ví dụ với Bảng 3.5 file huấn luyện sẽ như sau:

|  |
| --- |
| -1 1:1 2:1, 3:1 4:1 5:1 6:1  1 1:2 2:1 5:1 6:1 7:1 8:1 9:1 |

3.4. Tiền xử lý dữ liệu

Sau khi đã có dữ liệu em sẽ tiến hành tiền xử lý dữ liệu với 2 bước là tách từ vào lọc bộ từ điển.

***3.4.1. Tách từ***

Danh sách tập dữ liệu là các Status Tiếng Việt do vậy chúng ta cần phải tách từ trước khi xây dựng bộ từ điểm với mô hình n-gram.

Tiếng Việt có đặc điểm là từ có thể là từ đơn hoặc từ ghép vì thế khoảng trắng không còn là dấu hiệu phân cách từ (như tiếng Anh chẳng hạn). Việc phân tách một câu thành tập hợp đúng các từ có nghĩa là hết sức quan trọng đặc biệt đến kết quả dự đoán. Em xây dựng mô-đun tách từ bằng cách sử dụng thư viện vnTokenizer. Thư viện này được viết bằng JAVA với độ chính xác tách đúng từ theo công bố của tác giả là trong khoảng từ 96% đến 98%.



Hình 3.9: Quy trình tách từ.

Input: là một câu hoặc một văn bản được lưu dưới dạng tệp.

Output: là một chuỗi các đơn vị từ được tách.

Ví dụ sau đây minh họa kết quả của giai đoạn tách từ:

* Văn bản nguồn: “*Để có thể thực hiện rút trích tự động tóm tắt cũng như phân lớp văn bản với máy học vectơ hỗ trợ thì văn bản cần được biểu diễn dưới dạng thích hợp*”.
* Văn bản sau giai đoạn tách từ: “*Để có\_thể thực\_hiện rút trích tự\_động tóm\_tắt cũng\_như phân lớp văn\_bản với máy học vectơ hỗ\_trợ thì văn\_bản cần được biểu\_diễn dưới dạng thích\_hợp*”.

Trong quá trình đưa file dữ liệu chạy qua vnTokenizer có một số Status không tách từ được sẽ bị loại bỏ. Danh sách Status sau khi chạy sẽ được lưu vào file csv có tên là **vn\_tokenizer\_status.csv.**



Hình 3.10: File vn\_tokenizer\_status.csv chứa danh sách Status sau khi chạy qua vnTokenizer.

***3.4.2. Lọc bộ từ điển***

Với một dữ liệu gồm nhiều Status thì danh sách bộ từ điểm sẽ rất lớn trong đó có nhiều từ không có ý nghĩa trong việc dự doán, làm chậm quá trình xử lý. Để giảm bớt bộ từ điển em sẽ loại bỏ các từ có số lần xuất hiện ít hơn 5 lần và những ký từ đơn như “a”, “!”, “#”… và thay thế các chữ số thành #digit. Bảng 3.9 thống kê số lượng danh sách từ điển tương ứng với các mô hinh n-gram.

Bảng 3.9: Thống kế số lượng từ của tập dữ liệu.

|  |  |
| --- | --- |
| **Từ điển** | **Tống số còn lại** |
| Tập từ điển unigram | 12,923 |
| Tập từ điển bigram | 370,663 |
| Tập từ điển trigram | 1,230,451 |
| Trung bình | 538,012 |

Sau khi đã có bộ từ điển em sẽ tìm trọng số tương ứng và tạo file định dạng Liblinear. Với mỗi bộ từ điển sẽ tạo ra 3 file với 3 trọng số tương ứng là số lần xuất hiện, TF-IDF và Binary. Tổng cộng có 9 file như sau:

Bảng 3.10: Danh sách các file theo định dạng liblinear.

|  |  |  |
| --- | --- | --- |
| **Số thứ tự** | **Tên file** | **Mô tả** |
| 1 | Unigram\_count.libsvm | Bộ từ điển unigram với trọng số xuất hiện của từ |
| 2 | Unigram\_tfidf.libsvm | Bộ từ điển unigram với trọng số TF-IDF |
| 3 | Unigram\_binary.libsvm | Bộ từ điển unigram với trọng số Binary |
| 4 | Bigram\_count.libsvm | Bộ từ điển bigram với trọng số xuất hiện của từ |
| 5 | Bigram\_tfidf.libsvm | Bộ từ điển bigram với trọng số TF-IDF |
| 6 | Bigram\_binary.libsvm | Bộ từ điển bigram với trọng số Binary |
| 7 | Trigram\_count.libsvm | Bộ từ điển trigram với trọng số xuất hiện của từ |
| 8 | Trigram\_tfidf.libsvm | Bộ từ điển trigram với trọng số TF-IDF |
| 9 | Trigram\_binary.libsvm | Bộ từ điển trigram với trọng số Binary |

3.5. Kết quả thực nghiệm

Chạy lần lượt 9 file trên máy tính có cấu hình:

* **Hệ điều hành:** Destop Windows 10
* **Vi xử lý:** Intel Core i5
* **Bộ nhớ RAM:** 16 GB
* **Môi trường:** Java 8.
* **Liblinear phiên bản 2.11**

Kết quả độ chính xác như bảng sau:

Bảng 3.11: Kết quả độ chính xác của tập dữ liệu theo từng Status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | 69.41% | 69.57% | **70.05%** | 69.68% |
| **Bigram** | **66.96%** | 67.19% | 69.78% | 67.98% |
| **Trigram** | 67.65% | 67.76% | 70.00% | 68.47% |
| **Trung bình** | 68.01% | 68.17% | 69.95% | 68.71% |

Hàng dọc đầu tiên là danh sách tập từ điển và hàng ngang đầu tiên là danh sách các trọng số tương ứng. Bảng 3.11 cho thấy độ chính xác cao nhất 70.05% với tập từ điển unigram và trọng số TF-IDF. Kết quả độ chính xác thấp nhất là 66.96% thuộc về tập từ điển bigram với trọng số lần xuất hiện của từ. Chênh lệch giữa độ chính xác cao nhất và thấp nhất là 3.09%. Trung bình độ chính xác 9 file là 68.71%.

Hình 3.11: Biểu đồ thể hiện kết quả theo trọng số.

Theo hình 3.11 ta thấy nếu xét theo trọng số thì TF-IDF cho kết quả tốt nhất trung bình là 69.95% rồi đến trọng số Binary là 68.17% và số lần xuất hiện là 68.01%.

Hình 3.12: Biểu đồ thể hiện kết quả theo tập từ điển.

Ngược lại nếu xét trên tập từ điển thì unigram cho kết quả tốt nhất trung bình là 69.68% rồi đế trigram là 68.47% và cuối cùng đến từ điển bigram là 67.98% như biểu đồ hình 3.12.

Kết quả ở Bảng 3.11 cho thấy độ chính xác của việc dự đoán giới tính của người dùng trên từng Status riêng rẽ nhau. Việc dự đoán trên toàn bộ Status của tùng người dùng sẽ cho kết quả như bảng sau:

Bảng 3.12: Kết quả độ chính xác của tập dữ liệu theo từng người dùng.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | **93.87%** | 90.89% | 77.83% | 87.53% |
| **Bigram** | 93.08% | 91.32% | 76.42% | 86.94% |
| **Trigram** | 92.38% | 91.85% | **75.99%** | 86.74% |
| **Trung bình** | 93.11% | 91.35% | 76.75% | 87.07% |

Bảng 3.12 cho thấy độ chính xác cao nhất 93.87% với tập từ điển unigram và trọng số lần xuất hiện. Kết quả độ chính xác thấp nhất là 75.99% thuộc về tập từ điển trigram với trọng số TF-IDF. Chênh lệch giữa độ chính xác cao nhất và thấp nhất là 17.88%. Trung bình độ chính xác 9 file là 87.07%.

Hình 3.13: Biểu đồ thể hiện kết quả theo trọng số của tập dữ liệu theo từng người dùng.

Theo hình 3.13 ta thấy nếu xét theo trọng thì độ lệch khác xa nhau trung bình là 4.87% trong đó trọng số lần xuất hiện của từ cho kết quả tốt nhất trung bình là 93.11% rồi đến trọng số Binary là 91.35% và thấp nhất là TF-IDF 76.75%.

Hình 3.14: Biểu đồ thể hiện kết quả theo tập từ điển của tập dữ liệu theo từng người dùng.

Nếu xét trên tập từ điển thì độ chênh lệch là khá nhỏ chỉ 0.62% trong đó unigram cho kết quả tốt nhất trung bình là 87.53% rồi đế bigram là 86.94% và cuối cùng đến từ điển bigram là 86.74% như biểu đồ hình 3.14.

Từ Bảng 3.x và Bảng 3.x cho thấy. Nếu dự đoán theo từng Status thì trọng số TF-IDF cho kết quả tốt nhất nhưng theo người dùng thì kết quả không phải là tốt nhất mà là trọng số Binary. Điều này cho thấy mực độ quan trọng của một từ với việc dự đoán theo từng Status phụ thuộc vào việc từ đó trong toàn tập dữ liệu hơn là trong Status đó. Còn với theo người dùng, với việc 1 người có nhiều Status mức độ quan trọng của từ trong tập bộ tập dữ liệu thấp vì từ xuất hiện gần như ở người dùng nào cũng có, việc dự đoán phụ thuộc vào số lượng sử dụng từ của từng người dùng.

Để đánh giá số lượng tập dữ liệu ảnh hưởng đến độ chính xác của dự đoán em sẽ chia tập dữ liệu gốc thành các tập nhỏ ngẫu nhiên với số lượng Status của một tập lần lượt là 10000, 50000, 100000, 150000. Với các bước thực hiện tương tự như tập dữ liệu ban đầu em thu được kết quả với phương pháp 10-fold Cross validation như sau:

Bảng 3.13: Kết quả độ chính xác của tập dữ liệu với 10,000 Status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | **61.57%** | 62.53% | 64.10% | 62.73% |
| **Bigram** | 61.66% | 61.96% | 64.15% | 62.59% |
| **Trigram** | 62.00% | 62.16% | **64.45%** | 62.87% |
| **Trung bình** | 61.74% | 66.22% | 64.23% | 62.73% |

Theo bảng 3.13 độ chính xác cao nhất là 64.45% của tập từ điển trigram với trọng số TF-IDF và thấp nhất là 61.57% là tập từ điển unigram với trọng số lần xuất hiện của từ, độ chênh lệch của hai độ chính xác là 2.88%. Độ chính xác trung bình của cả tập dữ liệu là 62.73%.

Bảng 3.14: Kết quả độ chính xác của tập dữ liệu với 50,000 Status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | 65.99% | 66.08% | 67.11% | 66.39% |
| **Bigram** | **64.77%** | **64.77%** | 67.35% | 65.63% |
| **Trigram** | 65.19% | 65.21% | **67.45%** | 65.95% |
| **Trung bình** | 65.32% | 65.35% | 67.30% | 65.99% |

Theo bảng 3.14 giống như tập dữ liệu 10,000 Status, độ chính xác cao nhất thuộc tập từ điển trigram với trọng số TF-IDF là 67.45% và thấp nhất là 64.77% của hai file với tập từ điển bigram với 2 trọng số lần xuất hiện và Binary, độ chênh lệch của hai độ chính xác là 2.68%. Độ chính xác trung bình của cả tập dữ liệu là 65.99%.

Bảng 3.15: Kết quả độ chính xác của tập dữ liệu với 100,000 Status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | 67.68% | 67.97% | 68.68% | 68.11% |
| **Bigram** | **65.90%** | 66.10% | 68.39% | 66.80% |
| **Trigram** | 66.43% | 66.64% | **68.72%** | 67.26% |
| **Trung bình** | 66.67% | 66.90% | 68.60% | 67.39% |

Theo bảng 3.15, độ chính xác cao nhất vẫn thuộc tập từ điển trigram với trọng số TF-IDF là 68.72% và thấp nhất là 65.90% của tập từ điển bigram với trọng số lần xuất hiện, độ chênh lệch của hai độ chính xác là 2.82%. Độ chính xác trung bình của cả tập dữ liệu là 67.39%.

Bảng 3.16: Kết quả độ chính xác của tập dữ liệu với 150,000 Status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Binary** | **Tf-Idf** | **Trung bình** |
| **Unigram** | 68.59% | 68.78% | 69.45% | 68.94% |
| **Bigram** | **66.51%** | 66.63% | 69.29% | 67.48% |
| **Trigram** | 67.13% | 67.24% | **69.58%** | 67.98% |
| **Trung bình** | 67.41% | 67.55% | 69.44% | 68.13% |

Theo bảng 3.16, độ chính xác cao nhất thuộc tập từ điển trigram với trọng số TF-IDF là 69.58% và thấp nhất là 66.51% của tập từ điển bigram với trọng số lần xuất hiện, độ chênh lệch của hai độ chính xác là 3.07%. Độ chính xác trung bình của cả tập dữ liệu là 68.13%.

Hình 3.15: Biểu đồ kết quả độ chính xác trung bình của từng tập dữ liệu.

Hình 3.15 cho thấy độ chính xác tỉ lệ thuận với số lượng dữ liệu Status. Số lượng càng lới thì độ chính xác càng cao. Chênh lệch giữa tập dữ liệu lớn và tập dữ nhỏ nhất 10,000 Status là 6.12%. Độ lệch trung bình của 5 tập dữ liệu là 1.53%.

3.6. Kết luận chương

Chương này đã đưa ra các tiêu chuẩn đánh giá và cá phương pháp thực nghiệm thực hiện trên tập dữ liệu thu thập được. Các giai đoạn tiền xử lý dữ liệu để xây dựng lên file để đánh giá. Cuối cùng là các kết quả thực nghiệm.

KẾT LUẬN

1. Kết quả đạt được

Luận văn tiến hành nghiên cứu giải quyết bài toán dự doán giới tính người dùng mạng xã hội dựa trên nội dung bài viết nói chung và thực nghiệm với mạng xã hội Facebook và nội dung bài viết là tiếng Việt dựa vào đặc trưng. Bài toán là nền tảng cho nhiều ứng dụng quan trọng để dự đoán giới tính người dùng nói riêng và các thông tin khác nói chung.

Những kết quả chính mà luận văn đạt được:

* Nghiên cứu và tìm hiểu về bài toán dự đoán giới tính, trình bày một số phương pháp dự đoán giới tính đã được nghiên cứu trước đó.
* Phân tích hai đặc điểm của nội dung bài viết tiếng Việt phục vụ cho quá trình tiền xử lý.
* Tìm hiểu và áp dụng các công cụ tiền xử lý dữ liệu đầu vào
* Nghiên cứu và tìm hiểu về thuật toán Support Vector Machine trên hai lớp và nhiều lớp.
* Xây dựng chương trình lấy nội dung bài viết của người dùng trên mạng xã hội Facebook.
* Xây dựng chương trình huấn luyện và kiểm thử với bộ dữ liệu lấy được.

2. Hạn chế

* Hạn chế số lượng và chất lượng của dữ liệu ảnh hưởng đến kết quả dự đoán.
* Luận văn tập trung lấy dữ liệu và dự đoán giới tính người dùng trên mạng xã hội Facebook chưa thực nghiệm trên các mạng xã hội khác như Twitter, Youtube…

3. Hướng phát triển

* Xây dựng bộ dữ liệu lớn hoàn chỉnh, phong phú ở các mạng xã hội khác nhau.
* Cải thiện hiệu xuất, tăng tốc độ xử lý dữ liệu với dữ liệu lớn.
* Xây dựng hệ thống hoàn chỉnh cho các dữ liệu người dùng trên mạng xã hội, blog, comment…

DANH MỤC TÀI LIỆU THAM KHẢO

**Tài liệu Tiếng Anh**

1. Do Viet Phuong and Tu Minh Phuong. “*Gender Prediction Using Browsing History”.* [KSE (1) 2013](http://dblp.uni-trier.de/db/conf/kse/kse2013-1.html#PhuongP13): 271-283.
2. Argamon, S., M. Koppel, J. Fine & A. R. Shimoni (2003). Gender, genre, and writing style in formal written texts. Text, 23(3).
3. Popescu, A. & G. Grefenstette (2010). Mining user home location and gender from Flickr tags. In Proc. of ICWSM-10, pp. 1873–1876.
4. Katja Filippova. User Demographics and Language in an Implicit Social Network
5. Claudia Peersman, Walter Daelemans, Leona Van Vaerenbergh. Predicting Age and Gender in Online Social Networks
6. RE Fan, KW Chang, CJ Hsieh, XR Wang, CJ Lin. "LIBLINEAR: A library for large linear classification". Journal of machine learning research 9 (Aug), 1871-1874
7. PENG Qiu-fang, LIU Yang – Research of gender prediciton based on SVM with E-commerce data. Available from:

<http://lxbwk.njournal.sdu.edu.cn/EN/abstract/abstract3503.shtml>

1. Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. Available from:

<https://academic.oup.com/biomet/article-abstract/62/1/207/220350/Mendenhall-s-studies-of-word-length-distribution>

1. De Vel, O., Anderson, A., Corney, M., Mohay, G. M. (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64.
2. Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August.

**Website tham khảo**

1. <https://developers.facebook.com>
2. <https://www.csie.ntu.edu.tw/~cjlin/liblinear/>
3. <http://restfb.com>
4. <http://mccormickml.com/2013/08/01/k-fold-cross-validation-with-matlab-code/>

PHỤ LỤC