**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

[MỞ ĐẦU 1](#_Toc483787327)

[1. Tính cấp thiết của đề tài 1](#_Toc483787328)

[2. Tổng quan về vấn đề nghiên cứu 1](#_Toc483787329)

[3. Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu 2](#_Toc483787330)

[4. Cấu trúc luận văn 3](#_Toc483787331)

[Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH 6](#_Toc483787332)

[1.1. Giới thiệu bài toán dự đoán giới tính. 6](#_Toc483787333)

[1.2. Các phương pháp dự đoán giới tính 7](#_Toc483787334)

[1.3. Các phương pháp dự đoán giới tính dựa trên các bài biết của người dùng 7](#_Toc483787335)

[*1.3.1.* *Dự đoán giới tính dựa trên nội dung bình luận trên Youtube* 7](#_Toc483787336)

[*1.3.2.* *Dự đoán giới tính sử dụng bài viết từ blog* 7](#_Toc483787337)

[*1.3.3.* *Dự đoán giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy* 8](#_Toc483787338)

[1.4. Kết luận chương 11](#_Toc483787339)

[Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI 12](#_Toc483787340)

[2.1. Phạm vi áp dụng 12](#_Toc483787341)

[2.2. Các đặc trưng sử dụng 12](#_Toc483787342)

[*2.2.1.* *Đặc trưng text* 12](#_Toc483787343)

[*2.2.2.* *Đặc trưng ký hiệu đặc biệt* 13](#_Toc483787344)

[2.3. Kỹ thuật học máy SVM 13](#_Toc483787345)

[*2.3.1.* *Giới thiệu kỹ thuật học máy SVM* 13](#_Toc483787346)

[*2.3.2.* *Áp dụng kỹ thuật học máy SVM vào dự đoán giới tính* 16](#_Toc483787347)

[2.4. Kết luận chương 16](#_Toc483787348)

[Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ 17](#_Toc483787349)

[3.1. Các tiêu chuẩn đánh giá 17](#_Toc483787350)

[3.2. Phương pháp thực nghiệm 19](#_Toc483787351)

[3.3. Thu thập và mô tả dữ liệu 26](#_Toc483787352)

[*3.3.1* *Thu thập dữ liệu* 26](#_Toc483787353)

[*3.3.2* *Mô tả dữ liệu đầu vào* 32](#_Toc483787354)

[3.4. Xử lý dữ liệu 33](#_Toc483787355)

[3.5. Kết quả thực nghiệm 33](#_Toc483787356)

[2.6. So sách kết quả với các nghiên cứu đã có 38](#_Toc483787357)

[3.7. Kết luận chương 39](#_Toc483787358)

[KẾT LUẬN 40](#_Toc483787359)

[1. Kết quả đạt được 40](#_Toc483787360)

[2. Hạn chế 40](#_Toc483787361)

[3. Hướng phát triển 40](#_Toc483787362)

[DANH MỤC TÀI LIỆU THAM KHẢO 42](#_Toc483787363)

[PHỤ LỤC 43](#_Toc483787364)

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 | |
|  | Bigram | Bigram | 1-gram và 2-gram | |
|  | Trigram | Trigram | 1-gram, 2-gram và 3-gram | |
|  | **API** | Application Programming Interface | Giao diện lập trình ứng dụng | |
|  | Status | Status | Bài đăng của người dùng trên mạng xã hội Facebook | |
|  | Tweet | Tweet | Bài đăng của người dùng trên mạng xã hội Twitter | |
|  | NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên | |
|  | Facebook | Facebook | Mạng xã hội Facebook | |
|  | Twitter | Twitter | Mạng xã hội Twitter | |
|  | Youtube | Youtube | Mạng xã hội Youtube | |

DANH MỤC CÁC BẢNG BIỂU

[Bảng 3.1: Danh sách tập văn bản D gồm 2 câu là C1 và C2 19](#_Toc483867574)

[Bảng 3.2: Danh sách từ điển unigram 20](#_Toc483867575)

[Bảng 3.3: Danh sách từ điển bigram 21](#_Toc483867576)

[Bảng 3.4: Danh sách từ điển trigram 22](#_Toc483867577)

[Bảng 3.5: Danh sách từ điển unigram với trọng số xuất hiện của từ 23](#_Toc483867578)

[Bảng 3.6: Danh sách từ điển unigram với trọng số TF-IDF 25](#_Toc483867579)

[Bảng 3.7: Danh sách từ điển unigram với trọng số Binary 26](#_Toc483867580)

DANH MỤC CÁC HÌNH VẼ

[Hình 1.6: Ví dụ về hồi quy tuyến tính 10](#_Toc483867733)

[Hình 1.7: Quá trình khớp 10](#_Toc483867734)

[Hình 2.x: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất. 14](#_Toc483867735)

[Hình 2.x: Minh họa bài toán phân 2 lớp bằng phương pháp SVM 15](#_Toc483867736)

[Hình 3.1: Hình ảnh minh họa 10-fold Cross validation. 18](#_Toc483867737)

[Hình 3.2: Graph API cho phép lấy thông tin của người dùng. 28](#_Toc483867738)

[Hình 3.3: Access\_token của người dùng trên Facebook 29](#_Toc483867739)

[Hình 3.4: Minh họa cách lấy danh sách Status trên Facebook. 30](#_Toc483867740)

[Hình 3.5: Tạo project để hỗ trợ lấy nhiều danh sách Status. 31](#_Toc483867741)

[Hình 3.6: Định dạng mỗi dòng trong file csv chứa status lấy được. 31](#_Toc483867742)

[Hình 3.7: File full\_status.csv chứa tất cả Status lấy được. 32](#_Toc483867743)

[Hình 3.7: Minh họa những Status cần phải loại bỏ đi. 32](#_Toc483867744)

MỞ ĐẦU

1. Tính cấp thiết của đề tài

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, em 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”.***

2. Tổng quan về vấn đề nghiên cứu

Phương pháp dự đoán giới tính đã được nghiên cứu trong một thời gian dài. Ở giai đoạn đầu, hầu hết các nghiên cứu về lĩnh vực này tập trung vào việc nghiên cứu phân tích các đoạn văn bản được tạo ra từ người dùng mạng để xác định giới tính người viết là nam hay nữ. Các phương pháp nghiên cứu chủ yếu dựa trên các đặc trưng chẳng hạn như từ vựng, cú pháp hoặc các đặc trưng về nội dung với dữ liệu được thu thập từ email, blog, ý kiến... Gần đây, do sự phát triển của Internet và các kênh truyền thông trực tuyến trong đó có sự phát triển của mạng xã hội phương pháp dự đoán giới tính đã được mở rộng hơn thay vì phân tích văn bản các blog, bài báo.

Trong luận văn này sẽ giải quyết các vấn đề của dự đoán giới tính người dùng mạng xã hội trên nội dung bài viết là các bài Status trên mạng xã hội. Luận văn sử dụng phương pháp học tập phổ biến SVM - Support Vector Machine để dự doán dựa trên dữ liệu thu thập được trên mạng xã hội phổ biến nhất hiện nay là Facebook.

3. Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu

***Mục đích nghiên cứu:***

* Tìm hiểu về bài toán dự đoán giới tính người dùng.
* Các phương pháp dự đoán giới tính đã có.
* Nghiên cứu các kỹ thuật xử lý ngôn ngữ tiếng Việt.
* Kỹ thuật học máy và học máy SVM.
* Áp dụng và thực nghiệm dự đoán giới tính người dụng dựa trên nội dung bài viết văn bản tiếng Việt.

***Đối tượng nghiên cứu:***

* Các bài viết của người dùng trên mạng xã hội ở Việt Nam ở dạng text chuẩn tiếng Việt, không chứa hình ảnh, am thanh, video...
* Tiền xử lý văn bản tiếng Việt.
* Phương pháp SVM trong phân loại.

***Phạm vi nghiên cứu:***

* Nghiên cứu tổng quan về xử lý ngôn ngữ tự nhiên và bài toán dự đoán giới tính.
* Sau đó, luận văn đi sâu tìm hiểu ứng dụng phương pháp học máy SVM trong việc dự đoán giới tính người dùng trên mạng xã hội Facebook.

***Phương pháp nghiên cứu:***

* Tìm hiểu các phương pháp dự đoán giới tính hiện này đang có để xác định những điểm mạnh và hạn chế của các phương pháp đó.
* Các đặc trưng của nội dung bài viết ảnh hưởng đến việc xác định giới tính người dùng mạng xã hội.
* Tìm hiểu các kỹ thuật học máy hiện nay, nhưng ưu nhược điểm của các kỹ thuật học máy đó trong việc dự đoán giới tính người dùng mạng xã hội.
* Tìm kiểu chi tiết kỹ thuật học máy SVM và ứng dụng hiện này của phương pháp này.
* Tìm hiểu về mạng xã hội Facebook. Xác định những yếu tốt ảnh hưởng đến giới tính người dùng và các cách để lấy dữ liệu để nghiên cứu và thực nghiệm.
* Tìm hiểu phương pháp phân loại văn bản – Text categorization sẽ được sử dụng để phân tích nội dung bài viết hỗ trợ dự đoán giới tính người dùng trên mạng xã hội Facebook.

4. Cấu trúc luận văn

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 BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH

***Giới thiệu chương:***

Giới thiệu về bài toán xác định giới tính và áp dụng để xác định giới tính người dùng trên các mạng xã hội Facebook. Phần này cũng đưa ra các phương pháp xác định giới tính đã có trong đó chú ý đến phương pháp dựa trên nội dung bài viết.

***Nội dung chương:***

* Giới thiệu bài toán dự đoán giới tính
* Các phương pháp dự đoán giới tính
* Các phương pháp xác định giới tính dựa trên các bài biết của người dùng

***Kết luận chương:***

Chương này đã giới thiệu tổng quan về bài toán dự đoán giới tính, các phương pháp dự đoán và một số phương pháp xác định giới tính đã được nghiên cứu hiện này là tiền đề để phát triền luận văn.

CHƯƠNG 2: KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG MÃ HỘI

***Giới thiệu chương:***

Trình bày tổng quan về phương pháp học máy, một số kỹ thuật đã và đang được sử dụng trong việc phân tích người dùng mạng mã hội hiện nay. Dựa vào những đặc trưng nội dung bài viết khác nhau, sử dụng phương pháp học máy SVM để dự đoán giới tính người dùng.

***Nội dung chương:***

* Chi tiết về phạm vi áp dụng.
* Đưa ra các đặc trưng sử dụng để dự đoán giới tính người dùng mạng xã hội.
* Trình bày chi tiết thuật toán Máy vector hỗ trợ SVM.

***Kết luận chương:***

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: THỬ NGHIỆM VÀ ĐÁNH GIÁ

***Giới thiệu chương:***

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 đầu từ các bài viết trên mạng xã hội đã biết nhãn giới tính, sử dụng thư viện Liblinear có hỗ trợ kỹ thuật học máy SVM. Sau đó đưa bộ dữ liệu vào huấn luyện và sử dụng để dự đoán với bộ dữ liệu chưa có nhãn, đưa ra tỉ lệ và độ chính xác của phương pháp dự đoán dựa trên nội dung bài viết. Đánh giá kết quả so sách với các phương pháp dự đoán khác.

***Nội dung chương:***

* Xây dựng bộ dữ liệu huấn luyện và thử nghiệm
* Giai đoạn dự đoán
* Kết luận và hướng phát triển

***Kết luận chương:***

Dữ liệu huấn luyện và kiểm thử được xây dựng. Sau đó tiến hành giai đoạn dự đoán và đánh giá kết quả.

Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH

1.1. Giới thiệu bài toán dự đoán giới tính.

Bài toán dự đoán giới tính là một trong những bài toán kinh điển trong lĩnh vực xử lý dữ liệu văn bản. Đây là vấn đề có vai trò quan trọng khi phải xử lý một số lượng lớn dữ liệu. 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.

Trong lĩnh vực khai phá dữ liệu, các phương pháp phân loại văn bản đã dựa trên những phương pháp quyết định như quyết định Bayes, cây quyết định, k-láng giềng gần nhất, mạng nơron... Những phương pháp này đã cho kết quả chấp nhận được và được sử dụng trong thực tế. Trong những năm gần đây, phương pháp phân loại sử dụng phương pháp SVM được quan tâm và sử dụng nhiều trong những lĩnh vực nhận dạng và phân loại. SVM là một họ các phương pháp dựa trên cơ sở các hàm nhân (kernel) để tối thiểu hóa rủi ro ước lượng. Các thử nghiệm thực tế cho thấy, phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy, ...). So sánh với các phương pháp phân loại khác, khả năng phân loại của SVM là tương đương hoặc tốt hơn đáng kể

Vấn đề phân loại văn bản tiếng Việt được nhiều cơ sở nghiên cứu trong cả nước quan tâm trong những năm gần đây. Một số công trình nghiên cứu cũng đạt được những kết quả khả quan. Các hướng tiếp cận bài toán phân loại văn bản đã được nghiên cứu bao gồm: hướng tiếp cận bài toán phân loại bằng lý thuyết đồ thị, cách tiếp cận sử dụng lý thuyết tập thô, cách tiếp cận thống kê, cách tiếp cận sử dụng phương pháp học không giám sát và đánh chỉ mục. Nhìn chung, những cách tiếp cận này đều cho kết quả chấp nhận được. Tuy vậy để đi đến những triển khai khả thi thì vẫn cần đẩy mạnh nghiên cứu trên hướng này. Một trong những khó khăn trong việc áp dụng những thuật toán phân loại văn bản vào tiếng Việt là xây dựng được tập hợp từ vựng của văn bản. Vấn đề này liên quan tới việc phân tách một câu thành các từ một cách chính xác. Văn bản được biểu diễn dưới dạng vector và được phân loại theo phương pháp SVM. Trong luận văn này, trước hết chúng tôi trình bày cơ sở của phương pháp SVM và các thuật toán giải bài toán quy hoạch toàn phương phát sinh từ phương pháp này. Phần tiếp theo đề cập tới bài toán phân loại văn bản trong biểu diễn vector. Chúng tôi nhấn mạnh vào khía cạnh tiền xử lý văn bản, trích chọn đặc trưng, biểu diễn văn bản, và phân tích sự phù hợp của phương pháp SVM áp dụng vào bài toán phân loại văn bản. Phần cuối là các kết quả thí nghiệm ứng dụng SVM vào phân loại văn bản tiếng Việt. Những thí nghiệm này nhằm kiểm chứng khả năng phân loại của SVM đối với văn bản tiếng Việt. Đồng thời xác định các tham số của SVM thích hợp cho các phân lớp xác định trong bài toán phân loại văn bản

1.2. Các phương pháp dự đoán giới tính

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*

Ở trong những năm trở về trước, Blog là một loại nhật ký, website cá nhân phổ biến 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, ci, i = 1, ..., m, wi là một vector trọng lượng <wi 1, ..., wi n>, trong đó n là kích thước của tập hợp tính năng. Mỗi wi j, được khởi tạo để 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, người ta thường nhìn thấy được giới tính của người dùng. Ví dụ, Bạn có thể biết giới tính người dùng phía sau twitter sau đây?

AS LONG AS YOU LOVE ME <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 **điểm giữa kỳ**. 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*=*θ*​0​​+*θ*​1​​*x*​1​​+*θ*​2​​*x*​2 trong đó y là giới tính phụ thuộc x1 (nội dung) và x2 (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 =*θ*​0​​+*θ*​1​​*x*​1​​+*θ*​2​​*x*​2​​ 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 *θ*​0​​, *θ*​1​​, ...để xác định phương trình tuyến tính.



Hình 1.6: Ví dụ về hồi quy tuyến tính

Một trong những vấn đề gặp phải trong khi chạy mô hình Hồi Quy Tuyến Tính chính là hiện tượng quá khớp (overfitting). Overfitting là vấn đề xảy ra khi mô hình ta tạo ra cố gắng quá mức để khớp với các mẫu trong tập huấn luyện. Mô hình tuy rằng khớp với các mẫu huấn luyện nhưng lại không thể hiện được xu hướng của dữ liệu dẫn đến việc mô hình chỉ đúng với các giá trị trong tập huấn luyện và sai hoàn toàn với các giá trị test.



Hình 1.7: Quá trình khớp

Vấn đề quá khớp thường xảy ra khi bộ dữ liệu twitter của ta có nhiều đặc trưng nhưng lại có ít mẫu dữ liệu. Ví dụ như chúng ta muốn tạo ra một mô hình có dạng đường thẳng tức là cần hai đặc trưng x1, x2 (đặc trưng tọa độ trong mặt phẳng) nhưng lại chỉ có một mẫu dữ liệu được biểu diễn thành một điểm. Để xác định đường thẳng cần ít nhất hai điểm và nếu chỉ có một điểm thì có vô số mô hình phù hợp với mẫu dữ liệu nhưng trong đó chỉ có một mô hình là thật sự đúng với thực tế.

1.4. Kết luận chương

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

2.1. Phạm vi áp dụng

2.2. Các đặc trưng sử dụng

*2.2.1. Đặc trưng text*

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”.

*2.2.2. Đặc trưng ký hiệu đặc biệt*

2.3. Kỹ thuật học máy SVM

*2.3.1. Giới thiệu 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.

***a). 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.x: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất.

***b). 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.

***c). 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 xi thì cần phải xác định xi được phân vào lớp +1 hay lớp -1



Hình 2.x: 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.

***d). 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 co giãn (scaling) 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.

*2.3.2. Áp dụng kỹ thuật học máy SVM vào dự đoán giới tính*

2.4. Kết luận chương

Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. 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.

3.2. Phương pháp thực nghiệm

***3.2.1. 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:

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

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

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:

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

Bảng 3.2: Danh sách từ điển unigram

**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:

|  |  |
| --- | --- |
| **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 |

Bảng 3.3: Danh sách từ điển bigram

**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:

| **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 |

Bảng 3.4: Danh sách từ điển trigram

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 trình bày 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 = {, , ..., }, 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***

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 C1, “con”, “mèo”, “ngồi”, “trên”, “chiếc” và “mũ” mỗi từ xuất hiện 1 lần. Trong 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}

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

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

***b). Trọng số TF-IDF***

TF-IDF viết tắt của Term Frequency – Inverse Document Frequency, là trọng số của một 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}

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

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

***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}

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

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

***3.2.2. 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 |

3.3. Thu thập và mô tả dữ liệu

*3.3.1. 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 là hơn 700.000 Status và đượclưu vào filefull\_status.csv**.**



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

File full\_status.csv hiện tại có rất nhiều Status cần loại bỏ ví dụ như: không phải tiếng việt, có nhiều ký tự, hình vẽ, quá ngắn, quá dài, trùng nhau… Chúng ta cần phải loại bỏ những Status như vậy:





Hình 3.7: 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**.

*3.3.2. Mô tả dữ liệu đầu vào*

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Người dùng** | | **Bài viết** | |
| **Số lượng** | **Tỉ lệ** | **Số lượng** | **Tỉ lệ** |
| **Nam** |  |  |  |  |
| **Nữ** |  |  |  |  |
|  |  |  |  |  |

Bảng 3.x: Thống kế danh sách Status theo người dùng và bài viết

Bảng 3.x là thống kế tập dữ liệu đầu vào theo người dùng và theo bài viết:

* Với thống kế theo 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 bài viết 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.

Hình 3.x: Biểu đồ tỉ lệ nam/nữ theo người dùng.

Hình 3.x: Biểu đồ tỉ lệ nam/nữ theo bài viết.

3.4. Tiền xử lý dữ liệu

Sau khi đã có dữ liệu em sẽ tiến hành tiền xử lý dữ liệu:

* Bước 1: Tách từ.
* Bước 2: Xây dựng bộ từ điển.
* Bước 3: Đánh trọng số từ.
* Bước 4: Xây dựng file theo định dạng Liblinear.

3.4.1. Tách từ

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 [11]. 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.x: 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.x: File vn\_tokenizer\_status.csv chứa danh sách Status sau khi chạy qua vnTokenizer.

3.4.2. Xây dựng tập từ điển

Tiếp theo em sẽ xây dựng danh sách từ điển theo mô hình ngram. 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 chứ số thành #digit.

|  |  |
| --- | --- |
|  | **Số lượng** |
| Tập từ điển unigram |  |
| Tập từ điển bigram |  |
| Tập từ điển trigram |  |

3.4.3. Đánh trọng số từ

3.4.3. Xây dựng file theo định dạng Liblinear

3.5. Kết quả thực nghiệm

Bộ từ điển unigram gồm 3 file đánh giá:

* Unigram\_count.libsvm
* Unigram\_tfidf.libsvm
* Unigram\_binary-libsvm

|  |  |
| --- | --- |
|  | Độ chính xác |
| Unigram\_count.libsvm |  |
| Unigram\_tfidf.libsvm |  |
| Unigram\_binary.libsvm |  |
|  |  |

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 là 10000, 50000, 100000, 150000 và 200000.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram\_count.libsvm | Unigram\_tfidf.libsvm | Unigram\_binary.libsvm |
| 10000 |  |  |  |
| 50000 |  |  |  |
| 100000 |  |  |  |
| 150000 |  |  |  |
| 200000 |  |  |  |
|  |  |  |  |

Bộ từ điển bigram gồm 3 file đánh giá:

* Bigram\_count.libsvm
* Bigram\_tfidf.libsvm
* Bigram\_binary-libsvm

|  |  |
| --- | --- |
|  | Độ chính xác |
| Bigram\_count.libsvm |  |
| Bigram\_tfidf.libsvm |  |
| Bigram\_binary.libsvm |  |
|  |  |

Bộ từ điển trigram gồm 3 file đánh giá:

* Trigram\_count.libsvm
* Trigram\_tfidf.libsvm
* Trigram\_binary.libsvm

|  |  |
| --- | --- |
|  | Độ chính xác |
| Trigram\_count.libsvm |  |
| Trigram\_tfidf.libsvm |  |
| Trigram\_binary.libsvm |  |
|  |  |

2.6. So sách kết quả với các nghiên cứu đã có

3.7. Kết luận chương

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

**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