**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 tôi cùng sự giúp đỡ, khuyến khích của các quý thầy cô sau 2 năm tôi theo học chương trình đào tạo Thạc sĩ, chuyên ngành Khoa học máy tính của trường Học viện Công nghệ Bưu chính Viễn thông.

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. 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

Tôi 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 tôi trong quá trình học tập và nghiên cứu chương trình Thạc sĩ.

Tôi xin gửi lời cảm ơn sâu sắc tới **PGS.TS. Từ Minh Phương** đã tận tình hướng dẫn, giúp đỡ và động viên tôi 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. Tôi 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](#_Toc482570144)

[1. Tính cấp thiết của đề tài 1](#_Toc482570145)

[2. Tổng quan về vấn đề nghiên cứu 1](#_Toc482570146)

[3. Mục đích, đối tượng, phạm vi và phương pháp nghiên cứu 2](#_Toc482570147)

[4. Cấu trúc luận văn 3](#_Toc482570148)

[Chương 1 - GIỚI THIỆU BÀI TOÁN XÁC ĐỊNH GIỚI TÍNH 6](#_Toc482570149)

[1.1. Giới thiệu bài toán xác định giới tính. 6](#_Toc482570150)

[1.2. Các phương pháp xác định giới tính 8](#_Toc482570151)

[1.3. Các phương pháp xác định giới tính dựa trên các bài biết của người dùng 10](#_Toc482570152)

[*1.3.1.* *Dự đoán giới tính dựa trên nội dung bình luận trên Youtube* 10](#_Toc482570153)

[*1.3.2.* *Dự đoán giới tính sử dụng bài viết từ blog* 15](#_Toc482570154)

[*1.3.3.* *Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy* 15](#_Toc482570155)

[1.4. Kết luận chương 16](#_Toc482570156)

[Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG MÃ HỘI 18](#_Toc482570157)

[2.1. Phạm vi áp dụng 18](#_Toc482570160)

[2.2. Các đặc trưng sử dụng 30](#_Toc482570161)

[*2.2.1.* *Đặc trưng text* 30](#_Toc482570162)

[*2.2.2.* *Đặc trưng ký hiệu đặc biệt* 31](#_Toc482570163)

[2.3. Mô hình phân loại SVM 31](#_Toc482570164)

[*2.3.1.* *Giới thiệu kỹ thuật học máy SVM* 32](#_Toc482570165)

[*2.3.2.* *Áp dụng kỹ thuật học máy SVM vào dự đoán giới tính* 33](#_Toc482570166)

[2.4. Kết luận chương 35](#_Toc482570167)

[Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ 36](#_Toc482570168)

[3.1. Thu thập và mô tả dữ liệu đầu vào 36](#_Toc482570169)

[3.2. Các tiêu chuẩn dùng để đánh giá 40](#_Toc482570170)

[3.3. Phương pháp thực nghiệm 40](#_Toc482570171)

[3.4. Kết quả thực nghiệm 41](#_Toc482570172)

[2.5. So sách với một số phương pháp khác 41](#_Toc482570173)

[2.6. Độ phức tạp và thời gian thực hiện phương pháp 41](#_Toc482570174)

[3.7. Kết luận chương 42](#_Toc482570175)

[KẾT LUẬN 43](#_Toc482570176)

[1. Kết quả đạt được 43](#_Toc482570177)

[2. Hạn chế 43](#_Toc482570178)

[3. Hướng phát triển 44](#_Toc482570179)

[DANH MỤC TÀI LIỆU THAM KHẢO 45](#_Toc482570180)

[PHỤ LỤC 46](#_Toc482570181)

DANH MỤC TỪ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Ý nghĩa tiếng Anh** | **Ý nghĩa tiếng Việt** |
| 1 | SVM | Support vector machine | Máy vector hỗ trợ | |
| 2 | NB | Naïve Bayes |  | |
| 3 | kNN | K–Nearest Neighbor | 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 | RSS | Really Simple Syndication | Định dạng tài liệu | |

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Ngày này, 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 xác định 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 xác định rõ 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, tôi đã lựa chọn đề tài luận văn thạc sỹ là ***“Dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết”.***

2. Tổng quan về vấn đề nghiên cứu

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 xác định 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 xác định rõ 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.

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

***Mục đích nghiên cứu:***

* Tìm hiểu về bài toán xác định giới tính người dùng.
* Các phương pháp dự đoán giới tính đã có.
* Tìm hiểu các 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.

***Đối tượng nghiên cứu:***

* Nội dung bài viết của người dùng trên mạng xã hội ở Việt Nam.
* Bài toán tiền xử lý văn bản.
* Bài toán phân lớp.
* 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 XÁC ĐỊNH 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 hiện nay như: Facebook, Twitter, Youtube... 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 LibSVM 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 XÁC ĐỊNH GIỚI TÍNH

1.1. Giới thiệu bài toán xác định giới tính.

1.2. Các phương pháp xác định giới tính

1.3. Các phương pháp xác định giới tính dựa trên các bài biết của người dùng

*1.3.1. Dự đoán giới tính dựa trên nội dung bình luận trên Youtube*

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

*1.3.3. Xác định giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy*

1.4. Kết luận chương

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



2.1. Phạm vi áp dụng

2.2. Các đặc trưng sử dụng

*2.2.1. Đặc trưng text*

*2.2.2. Đặc trưng ký hiệu đặc biệt*

2.3. Mô hình phân loại SVM

*2.3.1. Giới thiệu kỹ thuật học máy SVM*

*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. Thu thập và mô tả dữ liệu đầu vào

3.2. Các tiêu chuẩn dùng để đánh giá

3.3. Phương pháp thực nghiệm

3.4. Kết quả thực nghiệm

2.5. So sách với một số phương pháp khác

2.6. Độ phức tạp và thời gian thực hiện phương pháp

3.7. Kết luận chương

KẾT LUẬN

1. Kết quả đạt được

Luận văn tiến hành nghiên cứu giải quyết bài toán phân loại văn bản tiếng Việt dựa vào đặc trưng. Bài toán là nền tảng cho nhiều ứng dụng quan trọng thực tế như lọc thư rác, hệ thống khuyến cáo/quan điểm người dùng…

Những kết quả chính mà đồ án đạt được:

* Nghiên cứu và tìm hiểu về bài toán phân lớp, trình bày một số thuật toán phân lớp tiêu biểu
* Tìm hiểu một số đặc điểm của văn bản 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 tin RSS hàng ngày
* Xây dựng chương trình huấn luyện và kiểm thử với bộ dữ liệu lấy từ chương trình lấy tin RSS.
* Xây dựng chương trình phân loại việc làm vào các category có sẵn từ các website không hỗ trợ RSS

2. Hạn chế

* Một văn bản đầu vào sau khi phân loại chỉ thuộc về một loại nhãn.
* Hạn chế số lượng và chất lượng của kho dữ liệu tin tức ảnh hưởng đến chất lượng phân loại của hệ thống.
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào.

3. Hướng phát triển

* Xây dựng bộ dữ liệu lớn hoàn chỉnh, phong phú về các lĩnh vực việc làm và chia các lĩnh vực càng nhỏ càng tốt
* Cần xác định giá trị chuẩn để một văn bản thuộc vào 1 hoặc nhiều thể loại, hoặc không thuộc thể loại nào
* Cải thiện hiệu xuất, tăng tốc độ xử lý dữ liệu
* Kiểm soát được thông tin lấy về theo ngày tháng năm mong muốn đảm bảo tính cập nhật của việc làm

DANH MỤC TÀI LIỆU THAM KHẢO

**Tài liệu Tiếng Việt**

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

**Tài liệu Tiếng Anh**

1. Abc
2. Abc

**Website tham khảo**

1. Abc

PHỤ LỤC