HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRƯƠNG CÔNG HẢI

ĐỀ CƯƠNG
LUẬN VĂN THẠC SĨ KỸ THUẬT

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

ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS. TỪ MINH PHƯƠNG

HÀ NỘI – 2016

I. MỞ ĐẦU

1. Lý do chọn đề tài

Ngày này, với sự phát triển của các mạng xã hội hiện nay trong đó có mạng mã hội Facebook có 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 cá nhân, tổ chức, công ty 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 đó là giới tính người dùng. Dựa vào một số nghiên cứu trước chúng ta có thể xác định giới tính người dùng dựa trên văn phong, cách dùng từ, diễn đạt trong các nội dùng bài viết và 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. Mục đích nghiên cứu

Mục tiêu nghiên cứu của đề tài là tìm hiểu và thử nghiệm được phương pháp dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết bằng cách sử dụng kỹ thuật học máy.

Bằng trực giác, ta có thể dự đoán một bài viết trên các mạng xã hội của người dùng có thể là nam hoặc nữ. Dựa vào những đặc trưng nội dung bài viết trên mạng xã hội và áp dụng kỹ thuật học máy trong đó có học máy SVM chúng ta có thể dự đoán giới tính người dùng. Dựa vào đó chúng ta có thể đưa ra các nhìn tổng quan hơn về tỷ lệ giới tính trên các mạng xã hội hiện này cũng là tiền đề cho các cá nhân, tổ chức, công ty đưa ra các đánh giá về hành vi người dùng, tỉ lệ chênh lệch giới tính, đưa ra chiến dịch quảng cáo sản phẩm phù hợp với từng giới tính người dùng trên mạng xã hội. Luận văn có áp dụng và thử nghiệm dự đoán giới tính người dùng trên các mạng xã hội hiên này. Mục tiêu cụ thể được trình bày trong luận văn như sau:

- Tìm hiểu về bài toán xác định giới tính người dùng.
- Các phương pháp xác định 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 dùng bài viết.

3. Đối tượng và phạm vi nghiên cứu

- Đối tượng: Nội dung bài viết của người dùng trên mạng xã hội
- Phạm vi nghiên cứu: Dự đoán giới tính người dùng trên mạng xã hội Facebook.

4. 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 phương pháp học máy đó trong việc dự đoán giới tính.
- Tìm kiểu chi tiết kỹ thuật học máy SVM và ứng dụng hiện này của phương pháp này
- Tìm hiểu các đặc trưng của 3 mạng xã hội trong phạm vi nghiên cứu. Xác định những yếu tốt ảnh hưởng đến giới tính người dùng và các cách để lấy dữ liệu để nghiên cứu
- Tìm hiểu phương pháp phân loại văn bản – Text categorization sẽ được sử dụng để phân tích comment hỗ trợ dự đoán giới tính người dùng trên mạng xã hội.

II. NỘI DUNG

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 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 1 sẽ bố cục theo các mục sau:

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

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 2 sẽ bố cục theo các mục sau:

2.1. Phạm vi áp dụng
2.2. Các đặc trưng sử dụng
  2.2.1. Đặc trưng text
  2.2.2. Đặc trưng các 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

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 dụng chương 3 sẽ bố cục theo các mục sau:

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
3.5. So sách với một số phương pháp khác
3.6. Độ phức tạp và thời gian thực hiện phương pháp
3.7. Kết luận chương

III. KẾT LUẬN

Áp dụng kỹ thuật học máy SVM trong việc dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết mang lại kết quả cao, phương pháp dự đoán đã kết hợp được nhiều yếu tố đặc biệt quan tâm đến nội dung ngôn ngữ của người dùng mạng xã hội từ đó xác định được giới tính đúng. Tuy nhiên phương pháp dự đoán vẫn còn một số hạn chế cần cải thiện trong tương lai và để xuất các hướng đi tiếp theo để nâng cao hơn nữa hiệu quả của việc dự đoán giới tình người dùng mạng mã hội.

IV. DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[01]. Do Viet Phuong and Tu Minh Phuong. "*Gender Prediction Using Browsing History".* KSE (1) 2013: 271-283.

[02]. Argamon, S., M. Koppel, J. Fine & A. R. Shimoni (2003). Gender, genre, and writing style in formal written texts. Text, 23(3).

[03]. Popescu, A. & G. Grefenstette (2010). Mining user home location and gender from Flickr tags. In Proc. of ICWSM-10, pp. 1873–1876.

[04]. Katja Filippova. User Demographics and Language in an Implicit Social Network

[05]. Claudia Peersman, Walter Daelemans, Leona Van Vaerenbergh. Predicting Age and Gender in Online Social Networks

V. DỰ KIẾN KẾ HOẠCH THỰC HIỆN

TT	Nội dung	Thời gian
1	Tìm hiểu các kỹ thuật học máy và học máy SVM.	10/2016 – 11/2016
2	Các cơ sở lý luận và các phương pháp dự đoán giới tính đã có	11/2016 – 12/2016
3	Các đặc trưng của nội dung bài viết sử dụng xác định giới tính người dùng	12/2016 – 01/2017
4	Cài đặt, thu thập bộ dữ liệu đầu vào	01/2017 – 03/2017
5	Thử nghiệm và đánh giá	04/2017

Ý KIẾN CỦA GIÁO VIÊN HƯỚNG DẪN	**NGƯỜI LẬP ĐỀ CƯƠNG**
(Ký ghi rõ họ tên)	*(Ký ghi rõ họ tên)*
PGS. TS. Từ Minh Phương	**Trương Công Hải**

DUYỆT CỦA TRƯỞNG TIỂU BAN CHẤM ĐỀ CƯƠNG

(Ký ghi rõ họ tên)