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

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – 2017

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS. TS. Từ Minh Phương**

Phản biện 1: .....……
.....…… ……....
.....…… ……....

Phản biện 2: …..……..
.....…… ……....
.....…… ……....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỤC LỤC

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".***

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

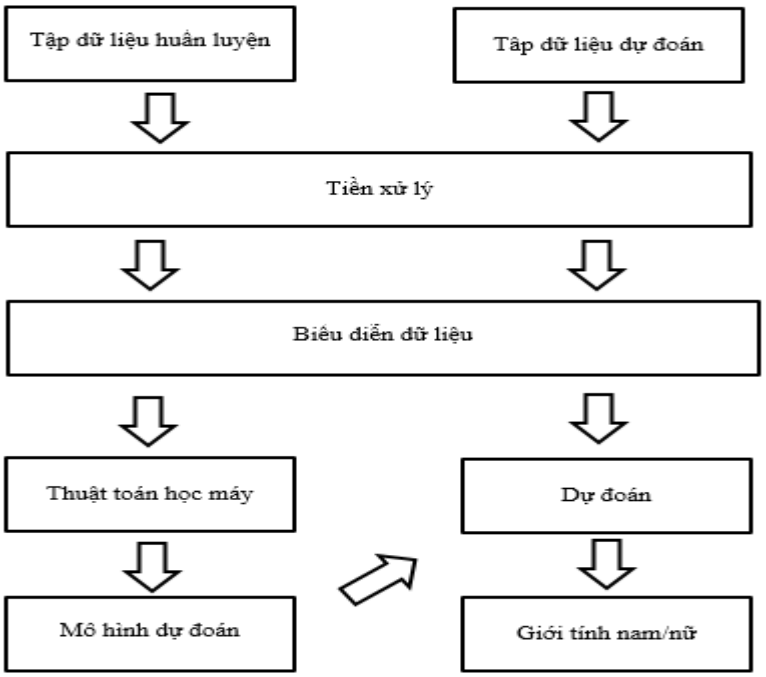
Ngày nay, với sự phát triển không ngừng của khoa học công nghệ cùng với sự hoàn thiện cơ sở hạ tầng và các trang thiết bị tương đối hiện đại và không ngừng phát triển. Theo báo cáo tổng kết của Bộ TT&TT năm 2016, tỷ lệ người sử dụng Internet ở Việt Nam đạt 62,76% dân số. Việc mọi người trao đổi thông tin liên lạc, tìm kiếm và cập nhật các thông tin về các lĩnh vực của mọi lĩnh vực tương đối dễ dàng và nhanh chóng.

Từ thực tế đó đã xuất hiện các nhu cầu muốn biết thông tin của người dùng Internet trong đó có thông tin giới tính. Trong nhiều trường hợp thông tin giới tính không có sẵn hoặc do họ không muốn người khác biết được khi đó xuất hiện bài toán dự đoán giới tính.

1.1.2. Bài toán dự đoán giới tính

Dự đoán giới tính (hay Determination Gender hoặc Gender Prediction) là quá trình phân loại và xác định giới tính Nam hoặc giới tính Nữ dựa trên dữ liệu đã biết trước.

Dưới đây là hình vẽ mô tả quy trình của bài toán dự đoán giới tính:



Hình 1.1: Quy trình bài toán dự đoán giới tính

Để tiến hành dự đoán giới tính nói chung, chúng ta sẽ thực hiện theo 2 phần chính là: Huấn luyện, Dự đoán

1.1.3. Ứng dụng của bài toán dự đoán giới tính

Hầu hết các thông tin đều là các hoạt động trực tuyến như tìm kiếm thông tin, chat, email, mua sắm trực tuyến... Từ đó việc dự đoán được thông tin người dùng trong đó có giới tính từ những dữ liệu đó sẽ giúp rất nhiều lợi ích như đưa ra các số liệu thống kê sử dụng theo giới tính người dùng, kế hoạch quảng cáo sản phẩm phù hợp với từng giới tính giúp giảm chi phí và tập trung hiệu quả hơn...

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.

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. [10] thực hiện năm 2007 với tập dữ liệu là tất cả blog được truy cập trong một ngày tháng 8 năm 2004.

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

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.

1.4. Kết luận chương

Chương này đã giới thiệu về bài toán dự đoán giới tính và ứng dụng, 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. Phạm vi bài toán

Trong luận văn tập trung vào bài toán dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết trên mạng xã hội Facebook. Dữ liệu bài viết trên Facebook chính là những bài đăng Status có nội dung văn bản đặc biệt của người dùng trên trang cá nhân. Chúng ta có thể chia thành 2 kiểu bài toán nhỏ:

- Dự đoán giới tính của người dùng với từng Status khác nhau.
- Dự đoán giới tính bằng cách kết hợp các Status của người dùng đó.

Luận văn sẽ tập trung vào việc dự đoán dựa trên các đặc trưng văn bản của nội dung bài viết Tiếng Việt cùng với việc áp dụng phương pháp học máy vector hỗ trợ SVM để dự đoán.

2.2. Đặc trưng văn bản và biểu diễn

2.2.1. 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.

2.2.2. Biểu diễn văn bản

Chúng ta cần biểu diễn văn bản một vector của các đặc trưng 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.

Ví dụ cho tập văn bản D gồm 2 câu C1 và C2 như Bảng 2.1:

Bảng 2.1: 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.

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

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 = $\{d_1, d_2, ..., d_m\}$ và T là một tập từ điển gồm n từ khác nhau T = $\{t_1, t_2, ..., t_n\}$.
- Output: Xây dựng w = (w_{ij}) là ma trận trọng số, trong đó w_{ij} là trọng số của từ $t_i \in$ T trong văn bản $d_j \in$ D.

a). Trọng số xuất hiện của từ (count)

Trọng số này được xác định bằng cách đếm số lần xuất hiện của từ $t_i \in$ T trong văn bản $d_j \in$ D.

$$w_{ij} = \text{số lần xuất hiện của từ } t_i \text{ trong văn bản } d_j.$$

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 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.

$$w_{ij} = TF - IDF(t_i, d_j, D)$$

Trọng số TF-IDF được tính như sau:

$$\text{TF-IDF}(t_i, d_j, D) = \text{TF}(t_i, d_j) \text{ x } \text{IDF}(t_i, D).$$

Trong đó:

$$\text{TF}(t_i, d_j) = \frac{\text{số lần từ } t_i \text{ xuất hiện trong văn bản } d_j}{tổng\ số\ từ\ trong\ văn\ bản\ d_j}$$

$$\text{IDF}(t_i, D) = \log\left(\frac{\text{Tổng số văn bản trong D}}{\text{Số văn bản có chứa từ } t_i,}\right)$$

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.

$$w_{ij} = \begin{cases} 1 & t_i \in d_j \\ 0 & t_i \notin d_j \end{cases}$$

2.3. Kỹ thuật học máy SVM

2.3.1. 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.

2.3.2. 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 $(x_1, y_1), (x_2, y_2),\ldots, (x_f, y_f)\}$ với $x_i \in R_n$, thuộc vào hai lớp nhãn: $y_i \in \{-1,1\}$ 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 $\overrightarrow{x_i}$ trong không gian: $\overrightarrow{x_i}.\vec{w} + b = 0$

$$\text{Đặt } f(\overrightarrow{X_i}) = sign(\overrightarrow{x_i}.\vec{w} + b) = \begin{cases} +1, & \overrightarrow{x_i}.\vec{w} + b > 0 \\ -1, & \overrightarrow{x_i}.\vec{w} + b < 0 \end{cases}$$

Như vậy, $f(X_i)$ biểu diễn sự phân lớp của X_i vào hai lớp như đã nêu. Ta nói $y_i = +1$ nếu X_i thuộc lớp I và $y_i = -1$ nếu X_i 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:

$$y_i(\sin(\overrightarrow{x_i}.\vec{w} + b)) \geq 1 \; với \; \forall \; i \in \overrightarrow{1,n}$$

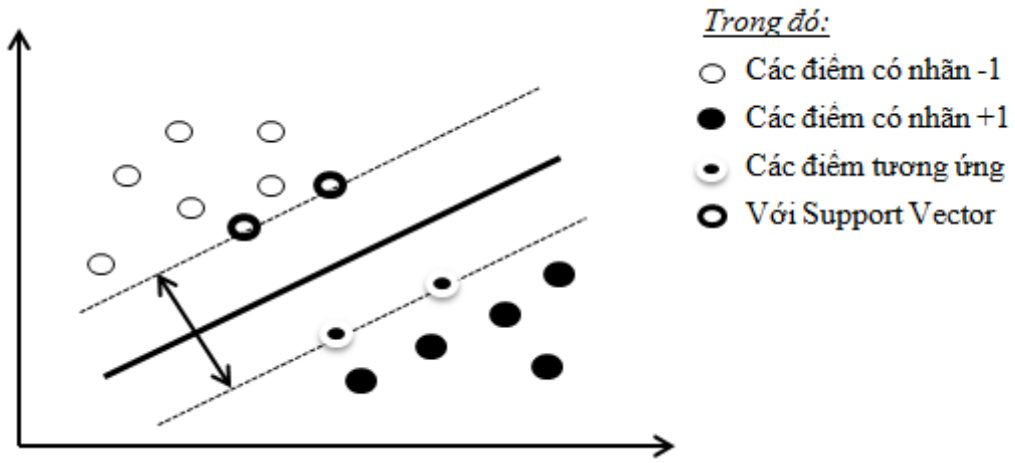
2.3.3. Bài toán phân 2 lớp với SVM

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 x_i thì cần phải xác định x_i 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 2.1: 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 $w \in \mathbb{R}^n$ là vector trọng số, $b \in \mathbb{R}^n$ là hệ số tự do, sao cho:

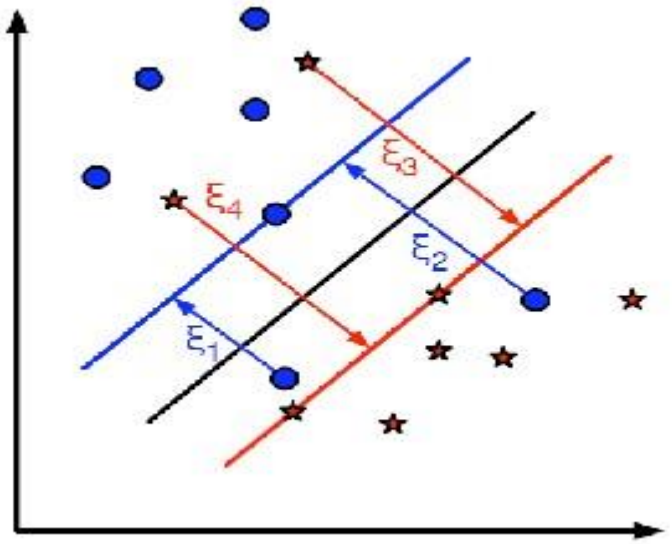
$$f(x_i) = sign(x_i.w^T + b) = \begin{cases} +1, y_i = +1 \\ -1, y_i = -1 \end{cases} \forall (x_i, y_i) \in D$$

Lúc này ta cần giải bài toán tối ưu:

$$\begin{cases} Min\big(L(w)\big) = \frac{1}{2}\|w\|^2 \\ y_i(x_i.w^T + b) \geq 1, i = 1, \dots, l \end{cases}$$

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 2.2: Tập dữ liệu được phân chia nhưng có nhiễu

Trong trường hợp này, ta sử dụng 1 biến mềm $\varepsilon_i \geq 0$ sao cho: $y_i(x_i.w^T + b) \geq$

$$1 - \varepsilon_i \, , i = 1, \dots, l$$

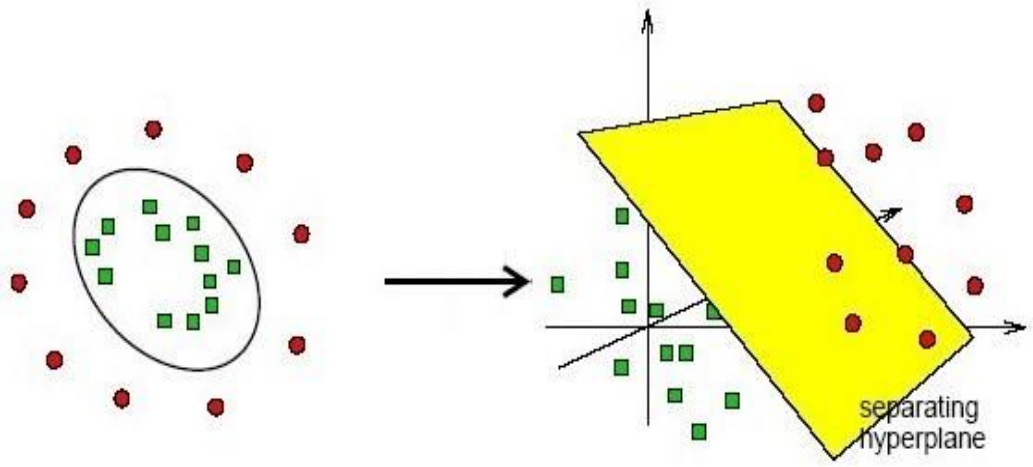
Bài toán tối ưu trở thành:

$$\begin{cases} Min\big(L(w,\varepsilon)\big) = \dfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \\ y_i(x_i.w^T + b) \geq 1 - \varepsilon_i, i = 1, \dots, l \\ \varepsilon_i \geq 0, i = 1, \dots, l \end{cases}$$

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 2.3: Tập dữ liệu không phân chia tuyến tính

Gọi ϕ là một ánh xạ phi tuyến từ không gian $\mathbb{R}^n$ vào không gian $\mathbb{R}^m$.

$$\phi: \mathbb{R}^n \to \mathbb{R}^m$$

Bài toán tối ưu trở thành:

$$\begin{cases} Min\big(L(w,\varepsilon)\big) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\varepsilon_i \\ y_i(\phi(x_i).w^T + b) \geq 1 - \varepsilon_i, i = 1, \dots, l \\ \varepsilon_i \geq 0, i = 1, \dots, l \end{cases}$$

2.3.4. Các bước chính của phương pháp 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.

2.3.5. Ưu điểm phương pháp SVM trong phân lớp dữ liệu

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 ước 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…).

2.4. 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.

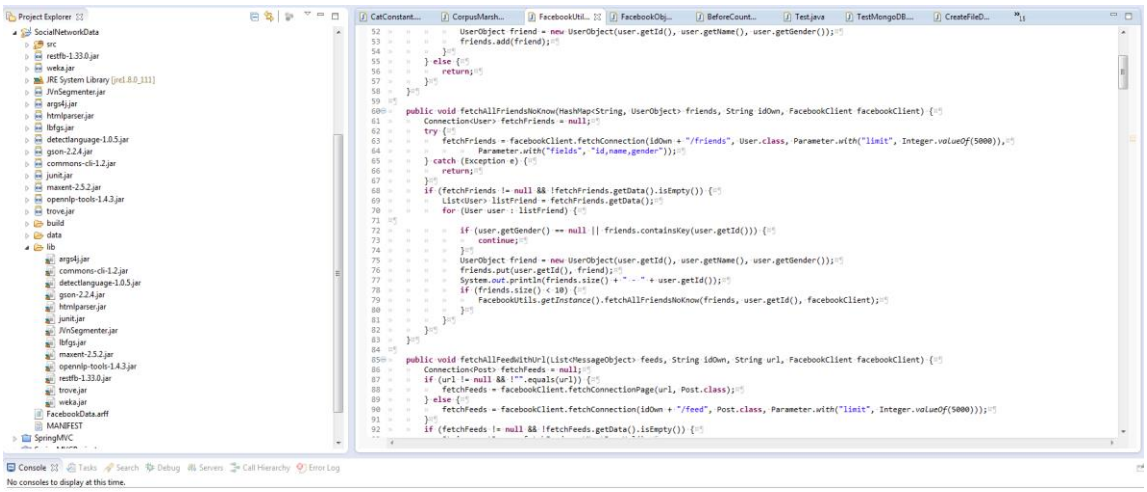
Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Thu thập và mô tả dữ liệu

3.1.1. Thu thập dữ liệu

Trên Facebook có cung cấp **Graph API** [15] cho phép lấy những thông tin người dùng trong đó có các bài Status của họ và bạn bè.

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 3.1: 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>

Số lượng Status lấy được lưu vào file **full_status.csv.**

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.
- Các Status không phải tiếng Việt.
- Những Status có quá nhiều kí tự hơn từ.

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 3.1 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 3.1: 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%

3.2. Các tiêu chuẩn đánh giá

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) = $\frac{TP+TN}{TP+TN+FP+FN}$.

Trong luận văn này sẽ sử dụng phương pháp **k-fold Cross validation** [18] với **10-fold** để thực hiện việc đánh giá.

3.3. Phương pháp thực nghiệm

Để tiến hành thực nghiệm với tập dữ liệu em sẽ sử dụng thư viện hỗ trợ phương pháp học máy SVM trong đó có bộ thư viện Liblinear [16]. 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. 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.

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.

Em xây dựng mô-đun tách từ bằng cách sử dụng thư viện vnTokenizer.

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.**

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 đoán, 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 3.2 thống kê số lượng danh sách từ điển tương ứng với các mô hinh n-gram.

Bảng 3.2: Thống kế số lượng từ của tập dữ liệu.

Từ điển	Tổng 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 3.3: 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

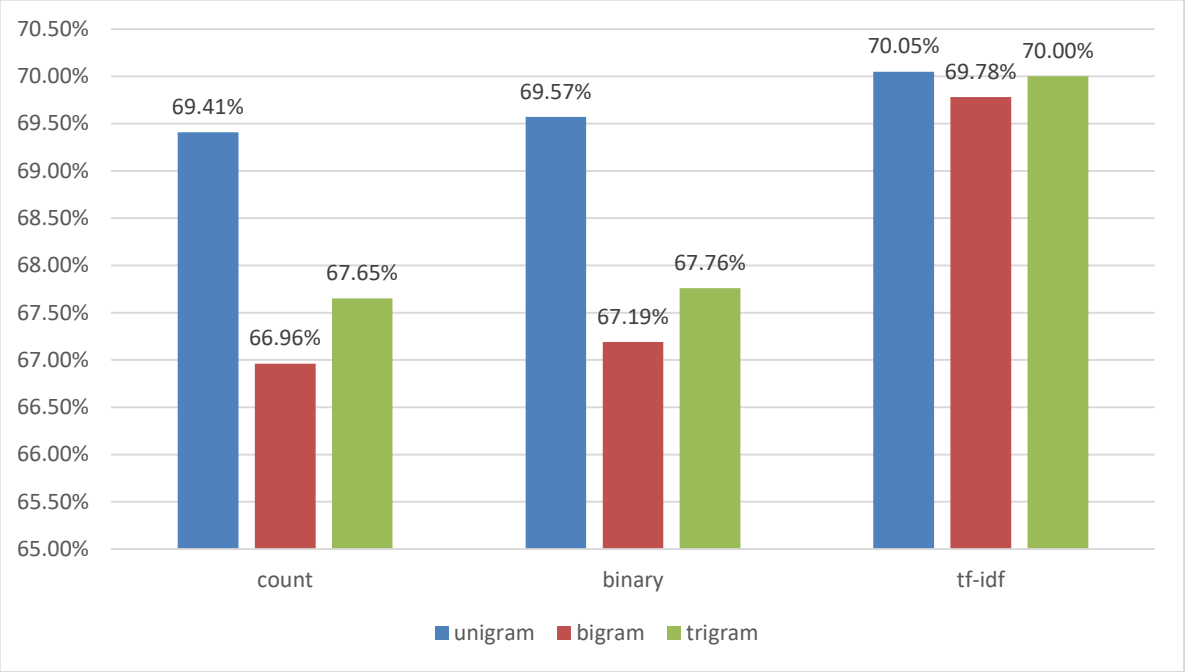
3	Unigram_binary.libsvm	Bộ từ điển unigram với trọng số Binary
4	Bigram_count.libsvm	Bộ từ điển bigram với trọng số xuất hiện của từ
5	Bigram_tfidf.libsvm	Bộ từ điển bigram với trọng số TF-IDF
6	Bigram_binary.libsvm	Bộ từ điển bigram với trọng số Binary
7	Trigram_count.libsvm	Bộ từ điển trigram với trọng 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

Bảng 3.4: Kết quả độ chính xác của tập dữ liệu theo từng Status.

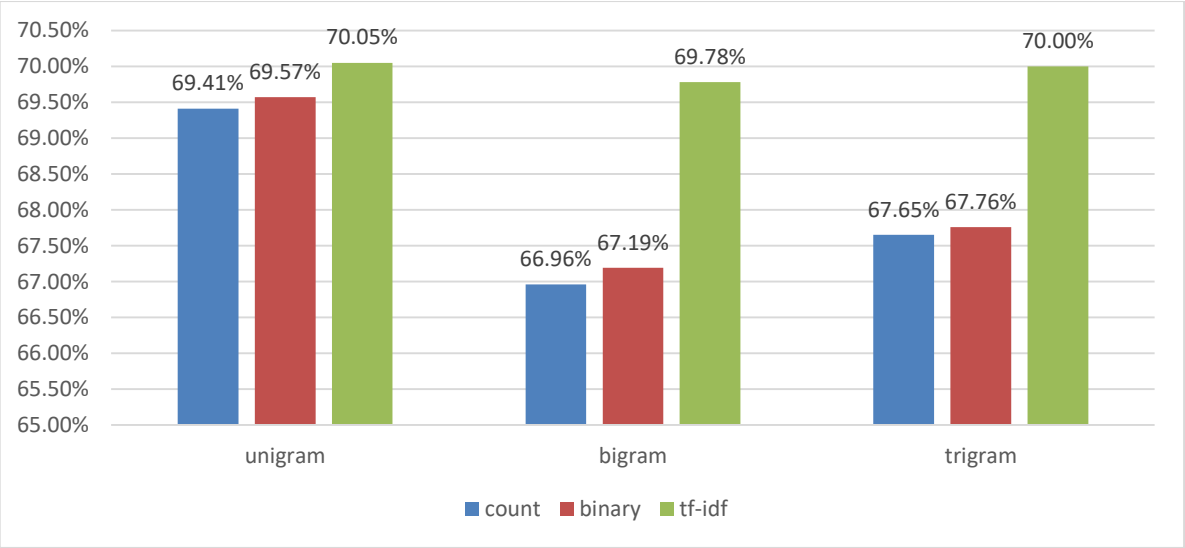
	Count	**Binary**	**Tf-Idf**	**Trung bình**
Unigram	69.41%	69.57%	**70.05%**	69.68%
Bigram	**66.96%**	67.19%	69.78%	67.98%
Trigram	67.65%	67.76%	70.00%	68.47%
Trung bình	68.01%	68.17%	69.95%	68.71%

Bảng 3.4 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 3.2: Biểu đồ thể hiện kết quả theo trọng số.

Theo hình 3.2 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 3.3: Biểu đồ thể hiện kết quả theo tập từ điển.

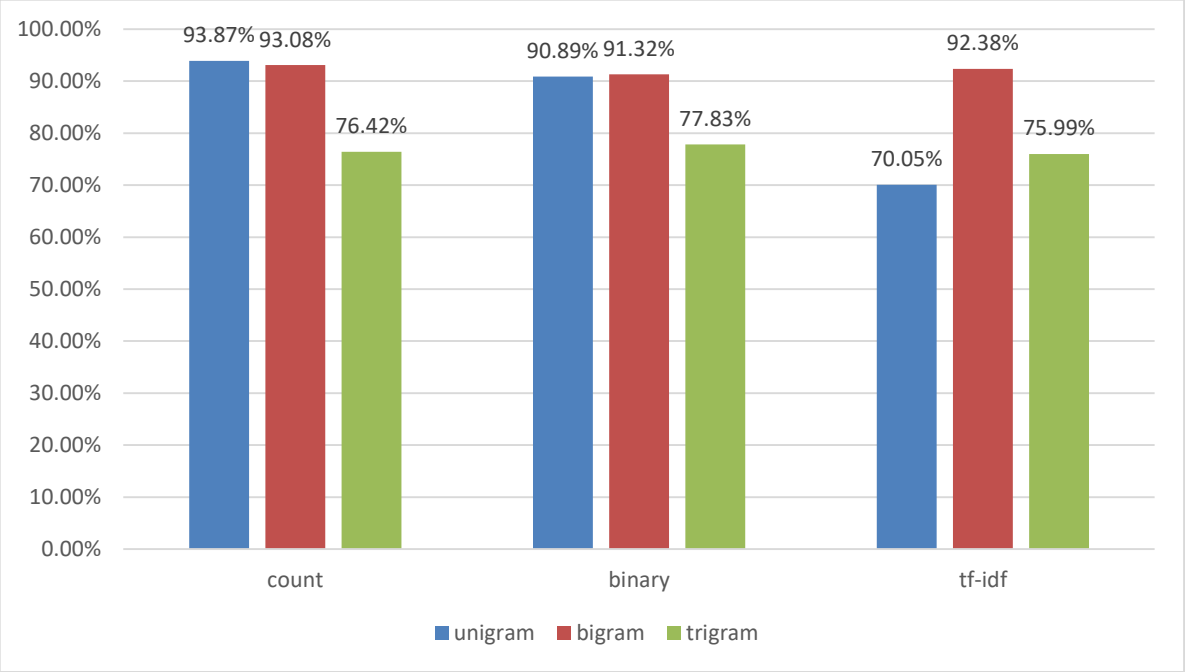
Ngược lại nếu xét trên tập từ điển thì unigram cho kết quả tốt nhất trung bình là 69.68% rồi đế trigram là 68.47% và cuối cùng đến từ điển bigram là 67.98% như biểu đồ hình 3.3.

Kết quả ở Bảng 3.4 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 3.5: Kết quả độ chính xác của tập dữ liệu theo từng người dùng.

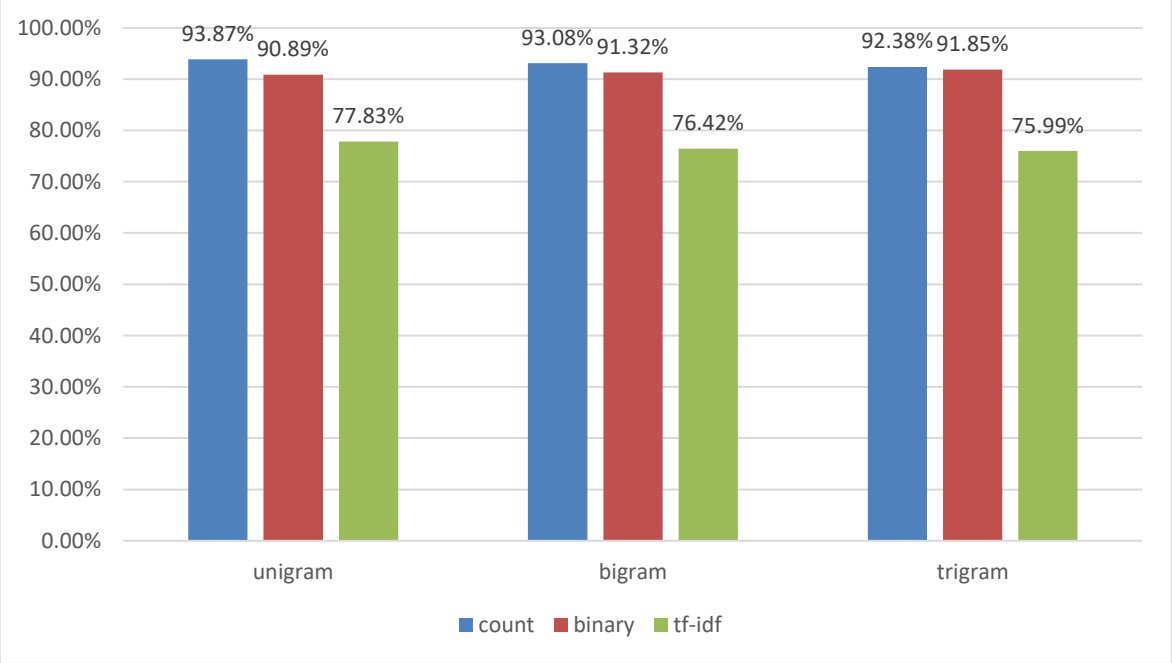
	Count	**Binary**	**Tf-Idf**	**Trung bình**
Unigram	**93.87%**	90.89%	77.83%	87.53%
Bigram	93.08%	91.32%	76.42%	86.94%
Trigram	92.38%	91.85%	**75.99%**	86.74%
Trung bình	93.11%	91.35%	76.75%	87.07%

Bảng 3.5 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 3.4: Biểu đồ thể hiện kết quả theo trọng số của tập dữ liệu theo từng người dùng.

Theo hình 3.4 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 3.5: 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 3.5.

Từ Bảng 3.4 và Bảng 3.5 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 3.6: 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%

Bảng 3.7: Kết quả độ chính xác của tập dữ liệu với 50,000 Status.

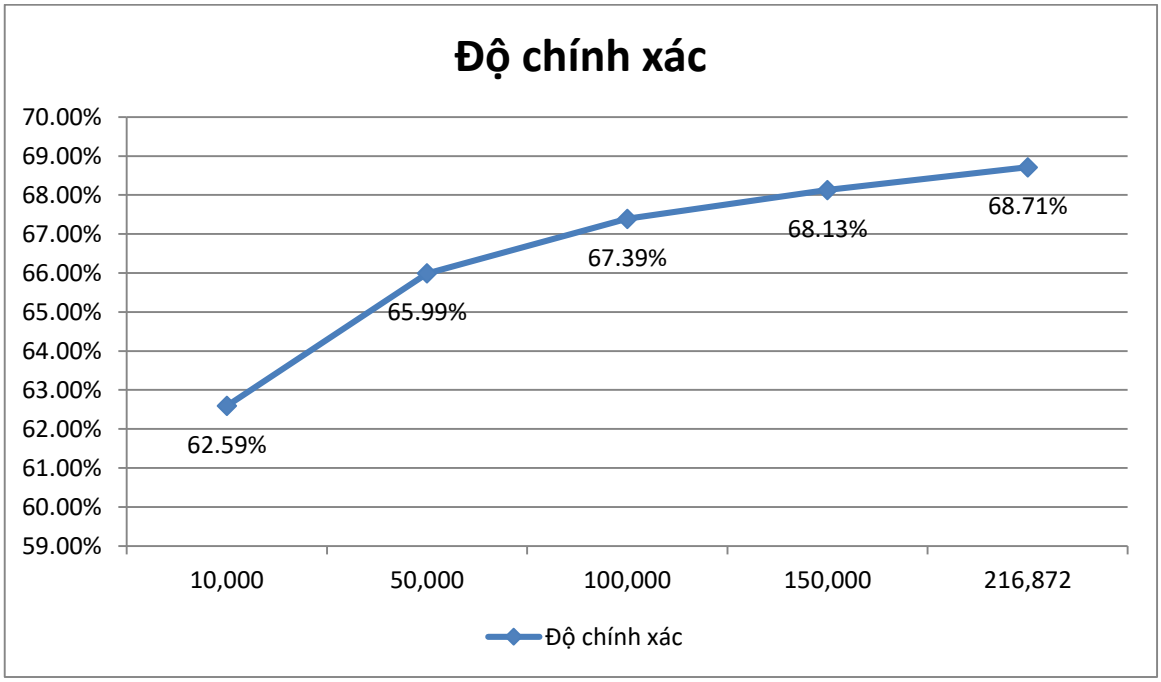
	Count	**Binary**	**Tf-Idf**	**Trung bình**
Unigram	65.99%	66.08%	67.11%	66.39%
Bigram	**64.77%**	**64.77%**	67.35%	65.63%
Trigram	65.19%	65.21%	**67.45%**	65.95%
Trung bình	65.32%	65.35%	67.30%	65.99%

Bảng 3.8: 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%

Bảng 3.9: 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%



Hình 3.6: Biểu đồ kết quả độ chính xác trung bình của từng tập dữ liệu.

Hình 3.6 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 suất, 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

[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.

[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

[06]. 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

[07]. 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

[08]. 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

[09]. 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.

[10]. Argamon, S., Koppel, M., Fine, J. and Shimoni, A. (2003). Gender, Genre, and Writing Style in Formal Written Texts, Text 23(3), August.

[11]. Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2008). Automatically Profiling the Author of an Anonymous Text, Communications of the ACM.

[12]. Burger, J. D., J. Henderson, G. Kim & G. Zarrella (2011). Discriminating gender on Twitter. In Proc. of EMNLP-11, pp. 1301–1309.

[13]. Nowson, S. & J. Oberlander (2006). The identity of bloggers: Openness and gender in personal weblogs. In Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, CA, 27-29 March 2006, pp. 163–167.

[14]. Yan, X. & L. Yan (2006). Gender classification of weblogs authors. In Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, Stanford, CA, 27-29 March 2006, pp. 228–230.

Website tham khảo

[15]. https://developers.facebook.com

[16]. https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[17]. http://restfb.com

[18]. http://mccormickml.com/2013/08/01/k-fold-cross-validation-with-matlab-code/