

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRƯỜNG 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 – 2017

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRƯỜNG 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 - 2017

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 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 CẢM ƠN

Em xin gửi lời cảm ơn và tri ân tới các thầy cô giáo, cán bộ của Học viện Công nghệ Bru 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 thầy **PGS. TS. Từ Minh Phương** đã tận tình hướng dẫn, giúp đỡ và động viên em để hoàn thành tốt nhất luận văn với đề tài là “**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 cảm ơn.

Tác giả

(Ký và ghi rõ họ tên)

Trương Công Hải

MỤC LỤC

MỞ ĐẦU	1
Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH.....	3
1.1. Giới thiệu bài toán dự đoán giới tính.	3
1.1.1. Mở đầu.....	3
1.1.2. Bài toán dự đoán giới tính.....	3
1.1.3. Ứng dụng của bài toán dự đoán giới tính.....	5
1.2. Các phương pháp dự đoán giới tính	5
1.3. Các phương pháp dự đoán giới tính dựa trên các bài viết của người dùng..	7
1.3.1. Dự đoán giới tính sử dụng bài viết từ blog.....	7
1.3.2. Dự đoán giới tính sử dụng dữ liệu từ các thông điệp trên twitter bằng phương pháp hồi quy	8
1.4. Kết luận chương	9
Chương 2 - KỸ THUẬT HỌC MÁY SVM VÀ ÁP DỤNG TRONG DỰ ĐOÁN GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI	10
2.1. Phạm vi bài toán	10
2.2. Đặc trưng văn bản và biểu diễn.....	11
2.2.1. Đặc trưng văn bản	11
2.2.2. Biểu diễn văn bản	12
2.3. Kỹ thuật học máy SVM.....	19
2.3.1. Ý tưởng.....	19

2.3.2.	<i>Cơ sở lý thuyết</i>	20
2.3.3.	<i>Bài toán phân 2 lớp với SVM</i>	21
2.3.4.	<i>Các bước chính của phương pháp SVM</i>	26
2.3.5.	<i>Ưu điểm phương pháp SVM trong phân lớp dữ liệu</i>	26
2.4.	<i>Kết luận chương</i>	27
Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ		28
3.1.	<i>Thu thập và mô tả dữ liệu</i>	28
3.1.1.	<i>Thu thập dữ liệu</i>	28
3.1.2.	<i>Mô tả dữ liệu đầu vào</i>	33
3.2.	<i>Các tiêu chuẩn đánh giá</i>	34
3.3.	<i>Phương pháp thực nghiệm</i>	35
3.4.	<i>Tiền xử lý dữ liệu</i>	36
3.4.1.	<i>Tách từ</i>	36
3.4.2.	<i>Lọc bộ từ điển</i>	38
3.5.	<i>Kết quả thực nghiệm</i>	39
3.6.	<i>Kết luận chương</i>	46
KẾT LUẬN		47
1.	<i>Kết quả đạt được</i>	47
2.	<i>Hạn chế</i>	47
3.	<i>Hướng phát triển</i>	47
DANH MỤC TÀI LIỆU THAM KHẢO		49

DANH MỤC TỪ VIẾT TẮT

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

DANH MỤC CÁC BẢNG BIỂU

<i>Bảng 2.1: Danh sách tập văn bản D gồm 2 câu là C1 và C2.....</i>	<i>12</i>
<i>Bảng 2.2: Danh sách từ điển unigram.....</i>	<i>13</i>
<i>Bảng 2.3: Danh sách từ điển bigram.....</i>	<i>13</i>
<i>Bảng 2.4: Danh sách từ điển trigram</i>	<i>14</i>
<i>Bảng 2.5: Danh sách từ điển unigram với trọng số xuất hiện của từ.....</i>	<i>16</i>
<i>Bảng 2.6: Danh sách từ điển unigram với trọng số TF-IDF.....</i>	<i>18</i>
<i>Bảng 2.7: Danh sách từ điển unigram với trọng số Binary.....</i>	<i>19</i>
<i>Bảng 3.1: Thống kê danh sách Status theo người dùng và bài viết.....</i>	<i>33</i>
<i>Bảng 3.2: Thống kê số lượng từ của tập dữ liệu.....</i>	<i>38</i>
<i>Bảng 3.3: Danh sách các file theo định dạng liblinear.....</i>	<i>38</i>
<i>Bảng 3.4: Kết quả độ chính xác của tập dữ liệu theo từng Status.....</i>	<i>39</i>
<i>Bảng 3.5: Kết quả độ chính xác của tập dữ liệu theo từng người dùng.....</i>	<i>41</i>
<i>Bảng 3.6: Kết quả độ chính xác của tập dữ liệu với 10,000 Status.....</i>	<i>44</i>
<i>Bảng 3.7: Kết quả độ chính xác của tập dữ liệu với 50,000 Status.....</i>	<i>44</i>
<i>Bảng 3.8: Kết quả độ chính xác của tập dữ liệu với 100,000 Status.....</i>	<i>44</i>
<i>Bảng 3.9: Kết quả độ chính xác của tập dữ liệu với 150,000 Status.....</i>	<i>45</i>

DANH MỤC CÁC HÌNH VẼ

<i>Hình 1.1: Quy trình bài toán dự đoán giới tính.....</i>	<i>4</i>
<i>Hình 1.2: Ví dụ về hồi quy tuyến tính</i>	<i>9</i>
<i>Hình 2.1: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất.....</i>	<i>20</i>
<i>Hình 2.2: Minh họa bài toán phân 2 lớp bằng phương pháp SVM</i>	<i>22</i>
<i>Hình 2.3: Tập dữ liệu được phân chia nhưng có nhiễu</i>	<i>23</i>
<i>Hình 2.4: Tập dữ liệu không phân chia tuyến tính</i>	<i>24</i>
<i>Hình 2.5: Ví dụ biểu diễn tập dữ liệu trên không gian 2 chiều.....</i>	<i>25</i>
<i>Hình 3.1: Graph API cho phép lấy thông tin của người dùng.....</i>	<i>28</i>
<i>Hình 3.2: Access_token của người dùng trên Facebook</i>	<i>29</i>
<i>Hình 3.3: Minh họa cách lấy danh sách Status trên Facebook.</i>	<i>30</i>
<i>Hình 3.4: Tạo project để hỗ trợ lấy nhiều danh sách Status.</i>	<i>31</i>
<i>Hình 3.5: Định dạng mỗi dòng trong file csv chứa status lấy được.</i>	<i>31</i>
<i>Hình 3.6: File full_status.csv chứa tất cả Status lấy được.</i>	<i>32</i>
<i>Hình 3.7: Minh họa những Status cần phải loại bỏ đi.....</i>	<i>33</i>
<i>Hình 3.8: Quy trình tách từ.....</i>	<i>37</i>
<i>Hình 3.9: File vn_tokenizer_status.csv chứa danh sách Status sau khi chạy qua vnTokenizer.</i>	<i>38</i>
<i>Hình 3.10: Biểu đồ thể hiện kết quả theo trọng số.</i>	<i>40</i>
<i>Hình 3.11: Biểu đồ thể hiện kết quả theo tập từ điển.</i>	<i>41</i>
<i>Hình 3.12: Biểu đồ thể hiện kết quả theo trọng số của tập dữ liệu theo từng người dùng.....</i>	<i>42</i>

Hình 3.13: Biểu đồ thể hiện kết quả theo tập từ điển của tập dữ liệu theo từng người dùng.....43

Hình 3.14: Biểu đồ kết quả độ chính xác trung bình của từng tập dữ liệu.....46

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 dung 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 thống kê, các kế hoạch quảng cáo cho 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 chung.

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

Với mục tiêu đặt ra luận văn sẽ được trình bày qua 3 chương như sau:

- ✚ Chương 1: Giới thiệu về bài toán dự đoán giới tính và ứng dụng thực tiễn. Phần này cũng đưa ra các phương pháp dự đoán giới tính đã có trong đó chú ý đến phương pháp dựa trên nội dung bài viết là tiền đề để phát triển luận văn.
- ✚ Chương 2: 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 này cũng trình bày chi tiết về kỹ thuật SVM là cơ sở lý thuyết để áp dụng vào thực hiện việc huấn luyện và dự đoán dựa trên nội dung bài viết trên mạng xã hội Facebook.

✚ Chương 3: Xây dựng các bước để thực nghiệm cho bài toán dự đoán giới tính người dùng mạng xã hội dựa trên nội dung bài viết. Lấy bộ dữ liệu từ các bài viết trên mạng xã hội Facebook, sử dụng thư viện Liblinear có hỗ trợ kỹ thuật học máy SVM. Sau đó đưa bộ dữ liệu vào xử lý và đánh giá kết quả thực nghiệm.

Chương 1 - GIỚI THIỆU BÀI TOÁN DỰ ĐOÁN GIỚI TÍNH

1.1. Giới thiệu bài toán dự đoán giới tính.

1.1.1. Mở đầu

Ngày nay, với sự phát triển không ngừng của khoa học công nghệ cùng với sự hoàn thiện cơ sở hạ tầng, 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 cần có bài toán dự đoán giới tính.

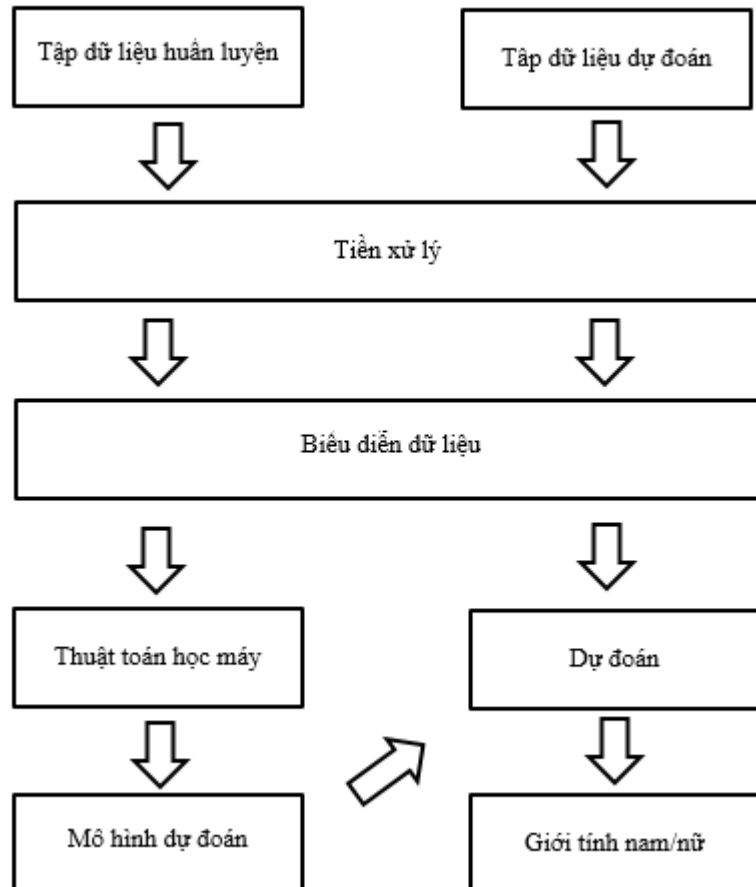
1.1.2. Bài toán dự đoán giới tính

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. Giống như những bài toán phân lớp đã được nghiên cứu trước đó. Dữ liệu để dự đoán giới tính rất đa dạng và phong phú đó có thể là một bài viết, lịch sử truy cập Internet, hình ảnh hoặc dữ liệu hành vi, thói quen... Trên thế giới đã có nhiều công trình nghiên cứu đạt những kết quả khả quan về việc dự đoán giới tính của người dùng.

Việc dự đoán có thể được tiến hành một cách thủ công như việc đọc nội dung của một câu, đoạn văn ta có thể đoán được giới tính của họ hay như quan sát các thông tin lịch sử... Tuy nhiên, đối với tập dữ liệu rất lớn trên mạng Internet thì phương pháp thủ công này sẽ tốn rất nhiều thời gian và công sức. Do vậy cần phải đưa ra phương pháp tự động để dự đoán giới tính. Phương pháp này giúp cho việc

dự đoán giới tính đạt độ chính xác cao với một dữ liệu rất lớn và sử dụng cho các mục đích như học tập, nghiên cứu, kinh doanh, tiếp thị thương mại...

Dưới đây là hình vẽ mô tả quy trình của bài toán 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:** Xây dựng mô hình dự đoán dựa trên tập dữ liệu thu thập của người dùng đã biết trước giới tính. Với tập dữ liệu huấn luyện sẽ đưa vào tiền xử lý sau đó được biểu diễn dữ liệu rồi sử dụng thuật toán học máy tạo ra mô hình dự đoán.

🚩 Dự đoán: Phần này sẽ đưa ra dự đoán với dữ liệu chưa biết giới tính. Dự vào mô hình dự đoán của phần huấn luyện. Dữ liệu cũng được tiền xử lý và biểu diễn như dữ liệu huấn luyện.

Đặc điểm nổi bật của bài toán là sự đa dạng về dữ liệu sử dụng để dự đoán nam và nữ giới. Các dữ liệu làm cho sự phân loại chỉ mang tính tương đối và có phần chủ quan, việc sử dụng, xử lý dữ liệu gì như thế nào tùy thuộc vào từng trường hợp. Ví dụ với người dùng đọc tin tức chúng ta có thể sử dụng dữ liệu lịch sử truy cập đọc các bài tin tức, còn trong trường hợp người dùng trên mạng xã hội chúng ta có thể dự đoán dựa trên những nội dung người dùng viết, bình luận...

1.1.3. Ứng dụng của bài toán dự đoán giới tính

Trên thế giới đã có một số công trình nghiên cứu với các hướng tiếp cận khác nhau cho bài toán dự đoán giới tính với các tập dữ liệu khác nhau. Các công trình tập trung vào việc dự đoán giới tính người dùng trên mạng Internet dựa trên những dữ liệu đã biết giới tính. Theo các kết quả trình bày trong các công trình đều cho kết quả khả quan.

Hiện nay, công nghệ ngày càng phát triển, đặc biệt với sự ra đời của các trang mạng xã hội, thương mại điện tử nên lượng thông tin, dữ liệu trao đổi lớn, phi cấu trúc, phức tạp, thậm chí là các thông tin rác cũng rất nhiều. Cần thiết phải có những nghiên cứu để xác định được thông tin gì là cần thiết và thông tin nào là dư thừa. Các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin đều đi tìm câu trả lời cho câu hỏi đó. Hầu hết các thông tin đều là các hoạt động trực tuyến như tìm kiếm thông tin, chat, email, mua sắm trực tuyến... Từ đó việc dự đoán được thông tin người dùng trong đó có giới tính từ những dữ liệu đã có 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. Như De Vel et al. [8] đã sử dụng 221 đặc trưng để xác định tác giả của email. Argamon và Koppel et al. [9] đã nghiên cứu sự khác biệt trong phong cách viết của nam và nữ trong 604 tài liệu của National Corpus của Anh. Schler et al. [10] khám phá việc sử dụng các đặc trưng và dựa trên nội dung để dự đoán giới tính và độ tuổi của các blogger trên bộ dữ liệu với hơn 71,000 bài viết blog từ blogger.com. mô hình đã đạt được kết quả chính xác là 80% cho dự đoán giới tính và 76% đối với các dự đoán tuổi. Nguyen et al. [7] đã tiến hành một nghiên cứu để dự đoán giới tính và độ tuổi của các thông điệp Twitter và diễn đàn bài viết bằng cách sử dụng phương pháp hồi quy với độ chính xác khoảng 80%.

Trong nghiên cứu của Burger et al. [12] đã sử dụng dữ liệu trên mạng xã hội Twitter để huấn luyện và dự đoán giới tính người dùng trên các tweet với từ và ký tự dựa trên đặc trưng n-gram đạt độ chính xác 75.5%. Khi thêm các đặc trưng tên đầy đủ, tên riêng của người dùng, độ chính xác đã tăng lên 89.1%, hơn nữa sử dụng thêm các bài viết mô tả về chính người dùng đã đạt được 92%. Ngoài ra, phương pháp tự huấn luyện khai phá dữ liệu không có nhãn được nghiên cứu nhưng hiệu suất thấp hơn.

Nghiên cứu của Nowson và Oberlander [13] đạt độ chính xác 92% trong việc dự đoán giới tính chỉ sử dụng đặc trưng n-gram. Dữ liệu của họ gồm 1,400/450 bài đăng viết bởi 47 nữ và 24 nam. Tuy nhiên, đặc trưng n-gram được chọn trước dựa vào việc từ đó có xuất hiện với tần số nhiều hay không trong ngôn ngữ của một giới trong giới khác. Khi tập dữ liệu hoàn chỉnh và được sử dụng để chọn các đặc trưng thì kết quả không đạt được độ chính xác như trước đó.

Yan et al. [14] đã sử dụng thuật toán phân loại Naïve Bayes để dự đoán giới tính của các tác giả blog. Trong tổng số 75,000 bài viết trên blog cá nhân của 3,000 tác giả với giới tính đã được ghi trên trang cá nhân. Kết quả thực nghiệm với độ chính xác 65%.

1.3. Các phương pháp dự đoán giới tính dựa trên các bài viết 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 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.

Nội dung nghiên cứu chú 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 tiền đề để giải quyết bài toán.

Nghiên cứu sử dụng mô hình MCRW (Multi-Class Real Winnow). Đối với mỗi lớp, c_i , $i = 1, \dots, m$, w_i một vector trọng lượng $\langle w_{i1}, \dots, w_{in} \rangle$, trong đó n là kích thước của tập thuộc tính. Mỗi w_{ij} được khởi tạo bắt đầu là 1. Các tập huấn luyện được sắp xếp ngẫu nhiên và được xử lý một lần. Thuật toán chạy vòng lặp huấn luyện liên tục, ngẫu nhiên đặt lại các ví dụ sau mỗi chu kỳ. Sau mỗi mười chu kỳ, Thuật toán kiểm tra số lượng các ví dụ đào tạo được phân loại chính xác. Nếu con số này đã giảm, thuật toán sẽ quay trở lại. Nếu không có cải tiến nào được tìm thấy sau năm vòng của 10 chu kỳ, thuật toán sẽ được chấm dứt.

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

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

a.) Giới thiệu

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

b.) Ý tưởng

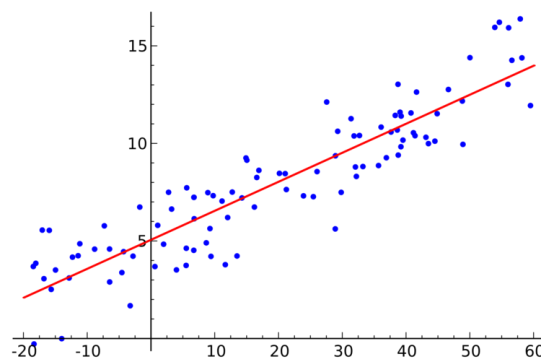
Đọc nội dung Twitter của ai đó, trong một số trường hợp, người ta có thể phân nào đoán được giới tính của người dùng. Ví dụ, Bạn có thể biết giới tính người dùng phía sau Twitter sau đây?

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

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

trung viết và các ký tự đặc biệt của 100 người dùng Twitter), ta xây dựng một phương trình $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ trong đó y là giới tính phụ thuộc x_1 (nội dung) và x_2 (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 $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ 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 $\theta_0, \theta_1 \dots$ để xác định phương trình tuyến tính.



Hình 1.2: Ví dụ về hồi quy tuyến tính

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 luận 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

Cùng với sự phát triển của Internet là việc hình thành và phát triển của các trang mạng xã hội trong đó có mạng xã hội Facebook đã trở thành công cụ thông tin liên lạc và chia sẻ cộng đồng phổ biến đối với hàng tỷ người trên thế giới không phân biệt không gian và thời gian. Facebook được thành lập bởi Mark Zuckerberg đây là một website mạng xã hội truy cập miễn phí, người dùng có thể tham gia để có thể kết bạn, chia sẻ, tìm kiếm thông tin, gửi tin nhắn và cập nhật trang hồ sơ cá nhân của mình để thông báo cho bạn bè biết... Trong đó có các Status (trạng thái) là đoạn nội dung của người dùng cá nhân cho phép họ thông báo cho bạn bè mọi người biết họ đang làm gì, ở đâu... trong các Status có thể là kết hợp văn bản, các ký hiệu đặc biệt, các hình ảnh, đường link hoặc các video để chia sẻ và cùng bàn luận.

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 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.

Tương tự như bài toán phân loại văn bản, đây là bài toán kinh điển trong lĩnh vực xử lý dữ liệu văn bản khi phải xử lý với một số lượng lớn dữ liệu nội dung trên Facebook. 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. 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. 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.

Về phân 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 nên: 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.

Về phân 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. Biểu diễn văn bản

Chúng ta cần biểu diễn văn bản thành một vector của các đặc trưng để dùng được giải thuật SVM 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 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.

Để có thể sử dụng được các thuật toán học máy cho văn bản, việc xây dựng tập từ điển để biểu diễn văn bản là rất quan trọng nó ảnh hưởng 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 2.1 gồm danh sách 9 từ như sau:

Bảng 2.2: Danh sách từ điển unigram

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

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

Ví dụ tập văn bản D ở Bảng 2.1 gồm danh sách 19 từ như sau:

Bảng 2.3: Danh sách từ điển bigram

Thứ tự	Từ
1	con
2	mèo
3	ngồi

4	trên
5	chiếc
6	mũ
7	chó
8	cắn
9	và
10	con mèo
11	mèo ngồi
12	ngồi trên
13	trên chiếc
14	chiếc mũ
15	con chó
16	chó cắn
17	cắn con
18	mèo và
19	và chiếc

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

Ví dụ tập văn bản D ở Bảng 2.1 gồm danh sách 29 từ như sau:

Bảng 2.4: Danh sách từ điển trigram

Thứ tự	Từ
1	con
2	mèo
3	ngồi
4	trên
5	chiếc
6	mũ
7	chó
8	cắn

9	và
10	con mèo
11	mèo ngồi
12	ngồi trên
13	trên chiếc
14	chiếc mũ
15	con chó
16	chó cắn
17	cắn con
18	mèo và
19	và chiếc
20	con mèo ngồi
21	mèo ngồi trên
22	ngồi trên chiếc
23	trên chiếc mũ
24	con chó cắn
25	chó cắn con
26	cắn con mèo
27	con mèo và
28	mèo và chiếc
29	và chiếc mũ

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

Bài toán

- ✚ Input: Cho một tập văn bản gồm m văn bản $D = \{d_1, d_2, \dots, d_m\}$ và T là một tập từ điển gồm n từ khác nhau $T = \{t_1, t_2, \dots, t_n\}$.
- ✚ 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} = số lần xuất hiện của từ t_i trong văn bản d_j .

Trong ví dụ ở Bảng 2.1 sử dụng tập từ điển Bảng 2.2. Trong câu C1, “con”, “mèo”, “ngồi”, “trên”, “chiếc” và “mũ” mỗi từ xuất hiện 1 lần. Trong câu C2, “con” xuất hiện 2 lần và “mèo”, “chiếc”, “mũ”, “chó”, “cắn” và “và” mỗi từ xuất hiện 1 lần. Như vậy trọng số cho C1 và C2 sẽ là:

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

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

Bảng 2.5: Danh sách từ điển unigram với trọng số xuất hiện của từ.

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

b). Trọng số TF-IDF

TF-IDF viết tắt của Term Frequency – Inverse Document Frequency, là trọng số của một từ thu được qua thống kê 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.

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

Trọng số TF-IDF được tính như sau:

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

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:

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

✚ 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 sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn. Ví dụ: từ nói (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:

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

Trong ví dụ ở Bảng 2.1 sử dụng tập từ điển Bảng 2.2. Trọng số TF-IDF cho từ “con” trong C1 được tính như sau:

$$\text{TF}(\text{“con”}, C1) = \frac{1}{6} = 0.1667$$

$$\text{IDF}(\text{“con”}, D) = \log\left(\frac{2}{2}\right) = 0$$

$$\text{TF-IDF}(\text{"con"}, C1, D) = \text{TF}(\text{"con"}, C1) \times \text{IDF}(\text{"con"}, D) = 0.1667 \times 0 = 0$$

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

$$\text{TF}(\text{"chó"}, C2) = \frac{1}{8} = 0.125$$

$$\text{IDF}(\text{"chó"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$

$$\text{TF-IDF}(\text{"chó"}, C2, D) = \text{TF}(\text{"chó"}, C2) \times \text{IDF}(\text{"chó"}, D) = 0.125 \times 0.301 = 0.038$$

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

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

Bảng 2.6: Danh sách từ điển unigram với trọng số TF-IDF.

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

c). Trọng số Binary

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

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

Trong ví dụ ở Bảng 2.1 sử dụng tập từ điển Bảng 2.2. Trong C1 có các từ “con”, “mèo”, “ngồi”, “trên”, “chiếc” và “mũ”. Trong C2 có các từ “con”, “mèo”, “chiếc”, “mũ”, “chó”, “cắn” và “và”. Như vậy trọng số cho C1 và C2 sẽ là:

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

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

Bảng 2.7: Danh sách từ điển unigram với trọng số Binary.

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

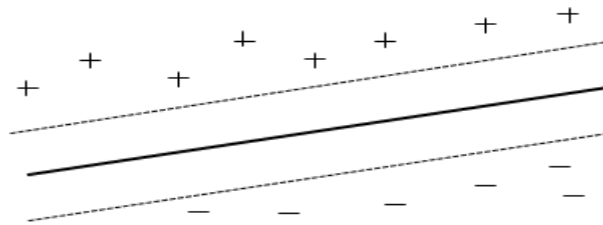
2.3. Kỹ thuật học máy SVM

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

2.3.1. 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.

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



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

2.3.2. 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 $(x_1, y_1), (x_2, y_2), \dots, (x_f, y_f)$ với $x_i \in \mathbb{R}_n$, thuộc vào hai lớp nhãn: $y_i \in \{-1, 1\}$ là nhãn lớp tương ứng của các x_i (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vectơ \vec{x}_i trong không gian: $\vec{x}_i \cdot \vec{w} + b = 0$

$$\text{Đặt } f(\vec{X}_i) = \text{sign}(\vec{x}_i \cdot \vec{w} + b) = \begin{cases} +1, & \vec{x}_i \cdot \vec{w} + b > 0 \\ -1, & \vec{x}_i \cdot \vec{w} + b < 0 \end{cases}$$

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:

$$y_i(\sin(\vec{x}_i \cdot \vec{w} + b)) \geq 1 \text{ với } \forall i \in \overline{1, n}$$

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 $\frac{1}{\|w\|}$. Cho dù các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

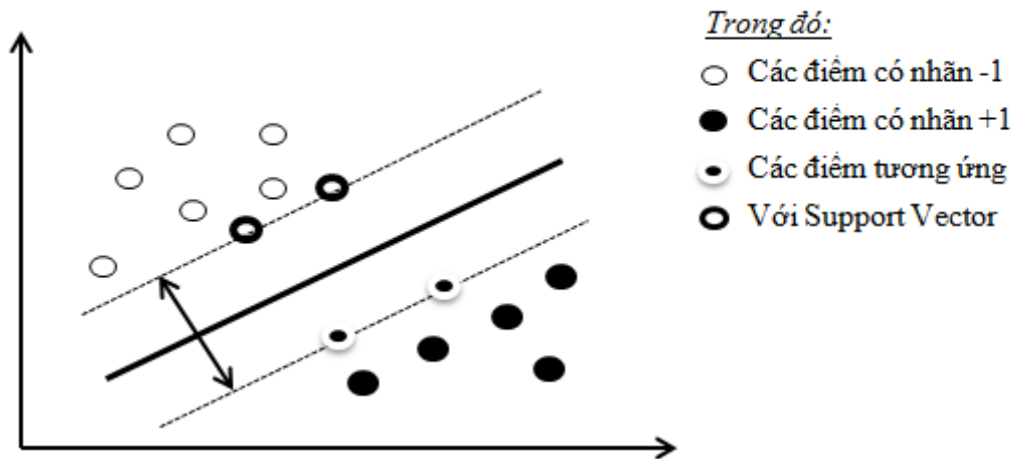
2.3.3. Bài toán phân 2 lớp với SVM

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 phẳng, tất cả các điểm được gán nhãn -1 thuộc về phía âm của siêu phẳng).



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

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách 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:

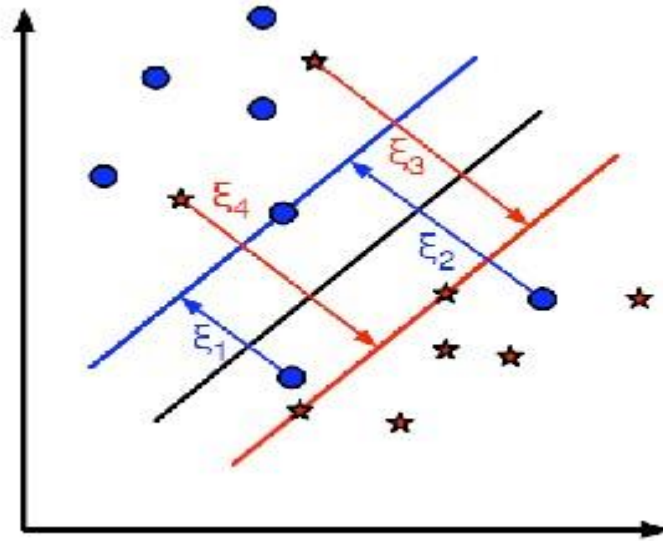
$$f(x_i) = \text{sign}(x_i \cdot w^T + b) = \begin{cases} +1, & y_i = +1 \\ -1, & y_i = -1 \end{cases} \forall (x_i, y_i) \in D$$

Lúc này ta cần giải bài toán tối ưu:

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

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 2.3: 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:

$$y_i(x_i \cdot w^T + b) \geq$$

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

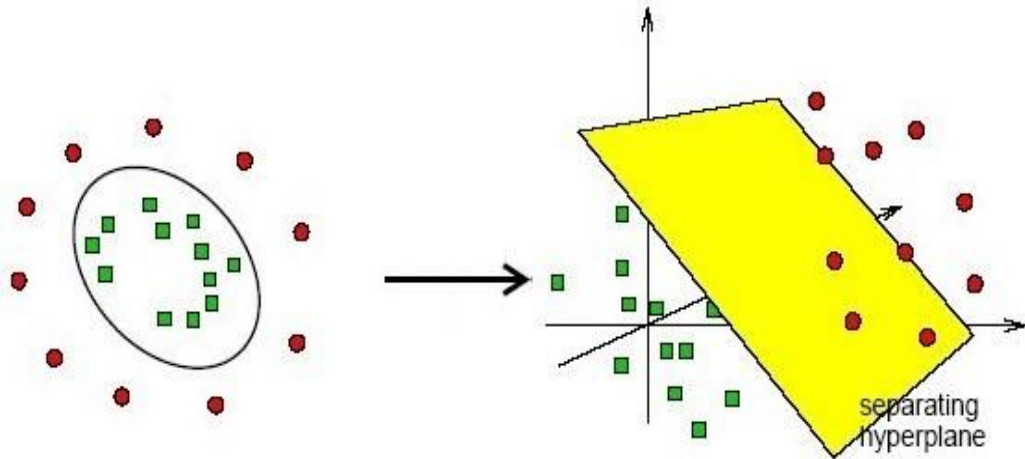
Bài toán tối ưu trở thành:

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

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 2.4: 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 .

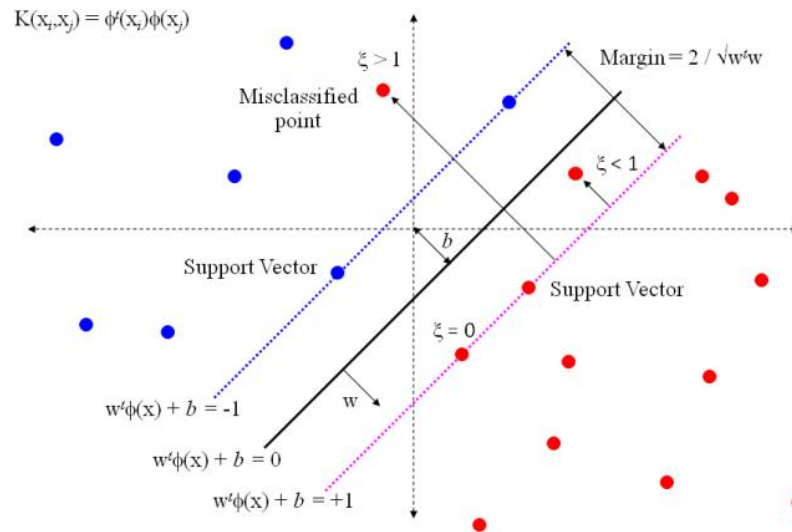
$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

Bài toán tối ưu trở thành:

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

Ví dụ:

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



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

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

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

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

2.3.4. Các bước chính của phương pháp SVM

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

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

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

Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng. 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

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

Chúng ta có thể thấy các thuật toán phân lớp hai lớp như SVM đều có đặc điểm chung là yêu cầu dữ liệu phải được biểu diễn dưới dạng vector đặc trưng, tuy nhiên các thuật toán khác đều phải sử dụng các 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...).

Trong công trình năm 1999, Joachims [11] đã so sánh SVM với Naïve Bayesian, k-Nearest Neighbour, Rocchio, và C4.5 và đến năm 2003, Joachims đã chứng minh rằng SVM làm việc rất tốt cùng với các đặc tính được đề cập trước đây của tập dữ liệu. Các kết quả cho thấy rằng SVM đưa ra độ chính xác phân lớp tốt nhất khi so sánh với các phương pháp khác.

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

2.4. 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 là bài toán phân 2 lớp của SVM là tiền đề để đánh giá với dữ liệu thực nghiệm.

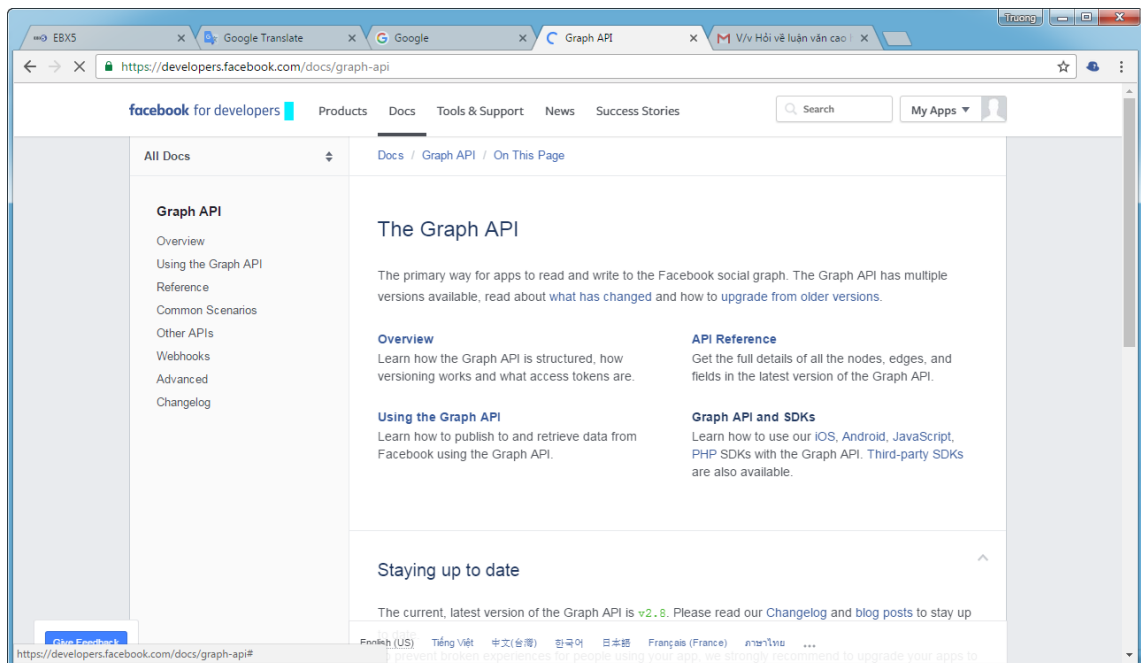
Chương 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Thu thập và mô tả dữ liệu

3.1.1. Thu thập dữ liệu

Dữ liệu sử dụng trong luận văn sẽ lấy các nội dung 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 dùng 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...

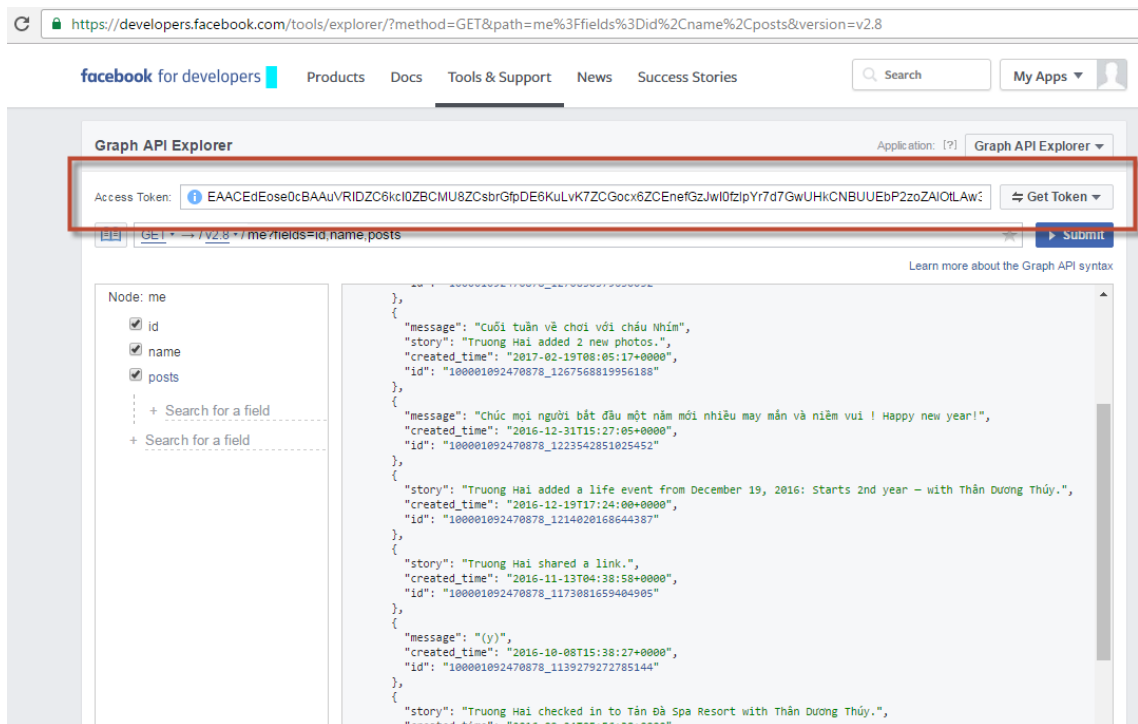
Trên Facebook có cung cấp **Graph API** [15] cho phép lấy những thông tin người dùng trong đó có các bài Status của họ và bạn bè.



Hình 3.1: 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**

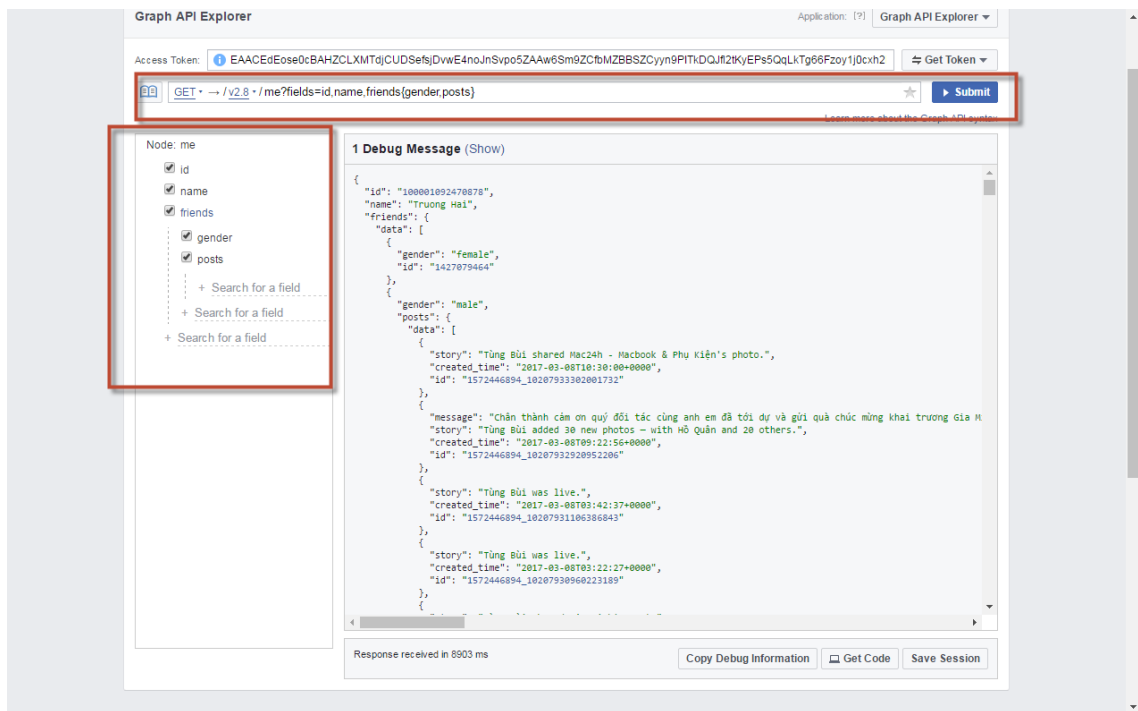
Access_token là của app trên facebook hoặc của tài khoản facebook.



Hình 3.2: Access_token của người dùng trên Facebook

Khi có access_token ta 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 các Status của 150 bạn bè.



Hình 3.3: Minh họa cách lấy danh sách Status trên Facebook.

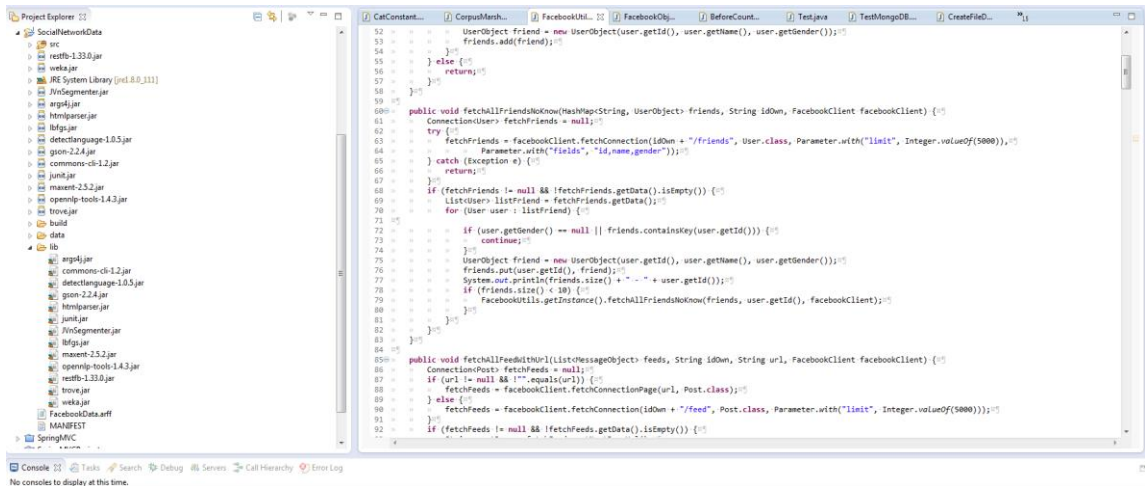
Để lấy được nhiều bài Status em có tạo 1 project java là **SocialNetworkData**

Trong project này em có sử dụng thư viện hỗ trợ là **restfb-1.33.0.jar** [17] 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

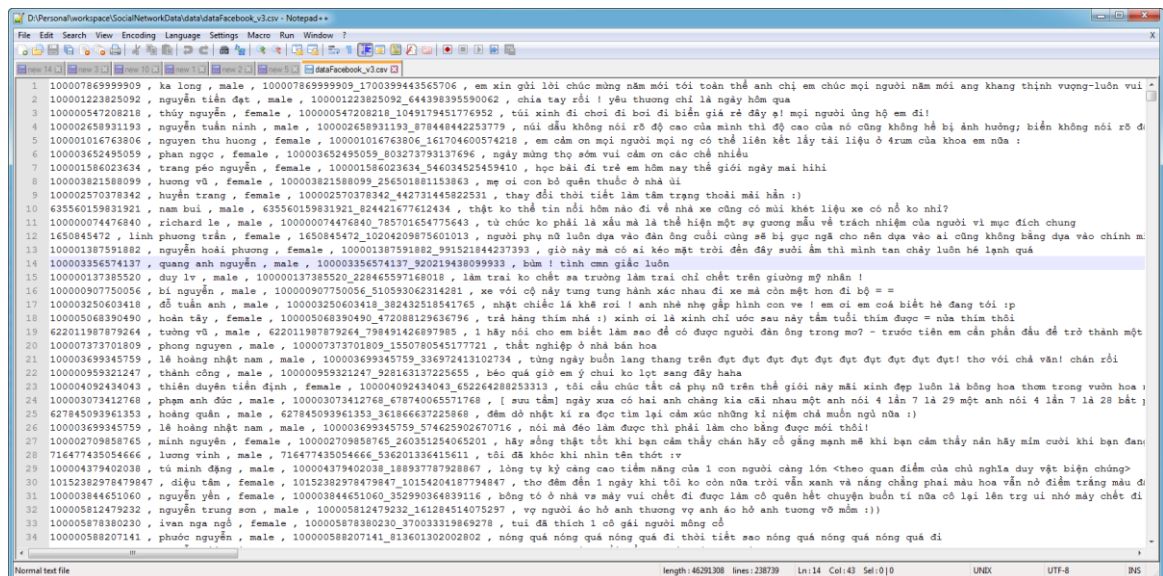
Để có sự đánh giá độ chính xác của phương pháp SVM em chỉ lấy dữ liệu người dùng đã có thông tin về giới tính rõ ràng (nam/nữ), chỉ lấy Status là văn bản thuần không chứa URL, tag bạn bè, hình ảnh, video...



Hình 3.4: 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>, < Id Status>, < Nội dung Status>



Hình 3.5: Định dạng mỗi dòng trong file csv chứa status lấy được.

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

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**.

Trong file **full_status_filter.csv** có chứa danh sách Status của nhiều người dùng khác nhau.

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

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

Nhận xét: Từ Bảng 3.1 thấy rằng nữ giới viết nhiều Status hơn nam giới. Trung bình một người dùng viết tầm 190 trong đó một nam giới là 163 Status và nữ giới có 226 Status.

3.2. Các tiêu chuẩn đánh giá

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

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

- ✚ Tập dữ liệu càng lớn thì độ chính xác càng tốt.
- ✚ Tập kiểm thử càng 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) = $\frac{TP+TN}{TP+TN+FP+FN}$.

Chỉ số Accuracy sẽ được sử dụng trong luận văn để 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 [18].

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

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

3.3. Phương pháp thực nghiệm

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

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


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

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

```
train -v 10 training_set_file
```

Trong đó:

 training_set_file: là file huấn luyện

 -v 10: có nghĩa là sử dụng 10-fold Cross validation

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

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

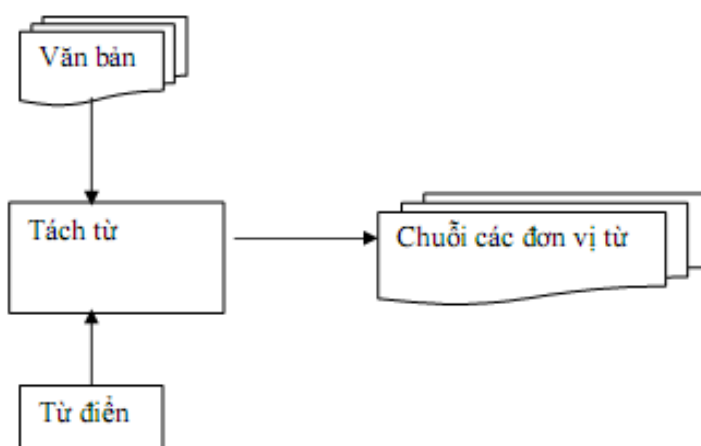
3.4. Tiền xử lý dữ liệu

Sau khi đã có dữ liệu em sẽ tiến hành tiền xử lý dữ liệu với 2 bước là tách từ và 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 điển với mô hình n-gram.

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



Hình 3.8: 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 vector 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_vector_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**.

```
100001098906226,Tu Pham,male,100001098906226_856834197696596,cái rap cạnh trường mình này : ) )
701749216554920,Lê Trung Hiếu,male,701749216554920_394624477267397,chuẩn bị đến cắt bả nào : ) )
100004103586449,Thư Nguyễn,female,100004103586449_437995769013873,chuối cả nhà có 1 thùng mới làm ăn phát đạt và vui_về cả tháng cả năm nhè : )
100002107197573,Hang Zozo,female,100002107197573_370746329672270,ngủ thôi mai về quê cho thoải mái ? híc
1572446894,Tùng Bùi,male,1572446894_1724869692877,dở biết mình ăn cái gì đó ?
100002434763938,Hiếu Vũ,female,100002434763938_821006674657150,cuộc sống có nhiều vấn đề mà
100005907964121,Po Linh'ss,female,100005907964121_402166009990328,dễ dàng sau mỗi bài hát con_gái hát - là một câu chuyện chưa kể ♥ dễ dàng sau mỗi lần con trai say -
702963976431168,Hằng Lê,female,702963976431168_487189071341994,từ lúc ngồi tiếp khách cho mẹ chả có ma nào vào nhà mình : v
100003846365578,Phuong Thủy,female,100003846365578_290856551052555,kết thúc rục rỏ các ty nhĩ @ @
100001098906226,Tu Pham,male,100001098906226_659047610808530,chưa năm nào ghét Tết như năm nay động đến cái quê gì cũng chò ra tết @ @
100001706443004,Dân Bách Phong,male,100001706443004_1202057733194430,chỉ có trong bóng tối tâm mới an được
100004640671859,Dương Nhất,female,100004640671859_482355798595772,thực ra đi ứng thời Tiết cũng tốt mỗi lần đi ứng là ngủ như mấy năm rồi không được ngủ ngủ tới
100002743926411,Hang Nguyen,female,100002743926411_424306687670792,sắp bị đuổi khỏi trường lại thấy yêu trường và tự hào về trường nhiều hơn mình nghĩ < 3
100001907330827,Côi Từ Bé,male,100001907330827_936411226432446,mong muốn chu nhất được nghỉ để về quê
10000086834288,Trần Việt Duy,male,10000086834288_769882013024686,tách rời lên 1 tấm cao mới let ' s go
100003508822141,Anh Tân,male,100003508822141_657023417757946,a e đi chơi tết thăm các cô chú trong làng
100002473612442,Lê Nam,male,100002473612442_1158455367580239,cuộc đời như thuốc phim dài !
702963976431168,Hằng Lê,female,702963976431168_1391338557593703,hằng nhà em mới về ai lấy gì ới em nha ; )
100003891892734,Khánh Ngọc,female,100003891892734_595837077222683,di học về chỉ phần hoa quả tráng miệng : v : v : v
100002874777572,Bình Nguyễn,female,100002874777572_408368165935672,sáng đây trời mưa cảm thấy thích thú lạ thường ta cầu cho 3ngày lễ sẽ mưa ! ! cho mọi người đ
100002107197573,Hang Zozo,female,100002107197573_708059999274233,cuộc sống làm cho con người ta xa nhau dần dần ghét cảm giác xa_lạ với những người mình quý_trợ
10000244434942,Phạm Thị Phương Thuý,female,10000244434942_513999801951496,chuẩn bị cho ngày mai ngày định mệnh : (
751209421613269,Trần Quyết Thắng,male,751209421613269_361081730626042,mắt tôi đã mờ đầu tôi đã đau người tôi đã mệt = > ngủ thôi
10000423224560,Đình Minh Bùi,male,10000423224560_168028120014928,* vì cuộc đời không có gì là mãi mãi * thế nên đừng chối_cải những điều xảy_ra trước_mắt bạn
676098852457280,Hoà Văn,female,676098852457280_852675048132992,cái này gọi là ngu thì chết tội tình gì : ) ) ~ ~
100002986113529,Dương Hoàng Lam Son,male,100002986113529_452404138202453,thông báo anh_em là dạo này uống thuốc không uống rượu nữa nhè : ( !
100001783978573,Trần Hải Đăng,male,100001783978573_930800890322701,có ai biết mua cái dây này ở đâu không nhỉ ? đang cần mua gấp ! @ @
100002107197573,Hang Zozo,female,100002107197573_224744410939130,khi vui hãy cười thật nhiều nhé vì như thế mình mới cảm thấy hạnh phúc ? hãy yêu hoàn_hảo 1 ngu
10000244434942,Phạm Thị Phương Thuý,female,10000244434942_477321332286010,sắp đến sinh nhật vk iu < 3
635560159831921,Nam Bui,male,635560159831921_736582906396312,hừ ta khinh của ngày_xưa đây ! giờ tên thật và đẹp nha lưu_ý là ảnh này mới thể_hiện đúng bản_chất
100003182677106,Tiểu Yết,female,100003182677106_503072596475492,hì vọng một năm mới với những khởi đầu mới thật thành công ! ! !
843301615734575,Thái Trang Anh Diệp,female,843301615734575_1151041404960593,miệt mài sx đã xong mẫu mới vài thoáng mắt mặc sang đẹp nhiều khách thích em sx liên
100004559322890,Thoa Nguyễn,female,100004559322890_413840582111249,lâu lắm hôm này lại dứt tay : ' (
676098852457280,Hoà Văn,female,676098852457280_545697732164060,loãng thoảng nhân ra lò mò hiểu được sợi dây đã đứt thì_có nói lại cũng vút đi mà thôi
100004559322890,Thoa Nguyễn,female,100004559322890_325398594288782,nhón nhà em ngập hết rồi : ( ( (
751209421613269,Trần Quyết Thắng,male,751209421613269_361081730626042,mắt tôi đã mờ đầu tôi đã đau người tôi đã mệt = > ngủ thôi
```

Hình 3.9: File vn_tokenizer_status.csv chứa danh sách Status sau khi chạy qua vnTokenizer.

3.4.2. Loại bỏ từ điển

Với một dữ liệu gồm nhiều Status thì danh sách bộ từ điển 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 hình n-gram.

Bảng 3.2: Thống kê 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

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

🚦 **Hệ điều hành:** Destop Windows 10

🚦 **Vi xử lý:** Intel Core i5

🚦 **Bộ nhớ RAM:** 16 GB

🚦 **Môi trường:** Java 8.

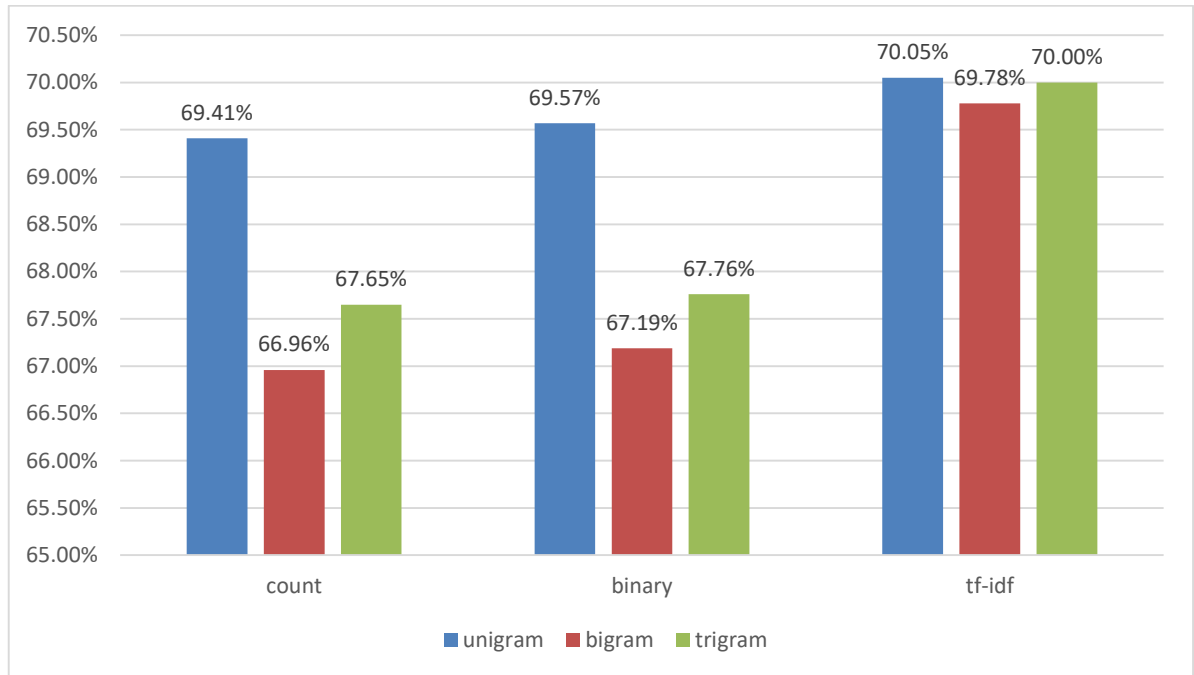
🚦 **Liblinear phiên bản 2.11**

Kết quả độ chính xác như Bảng 3.4:

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

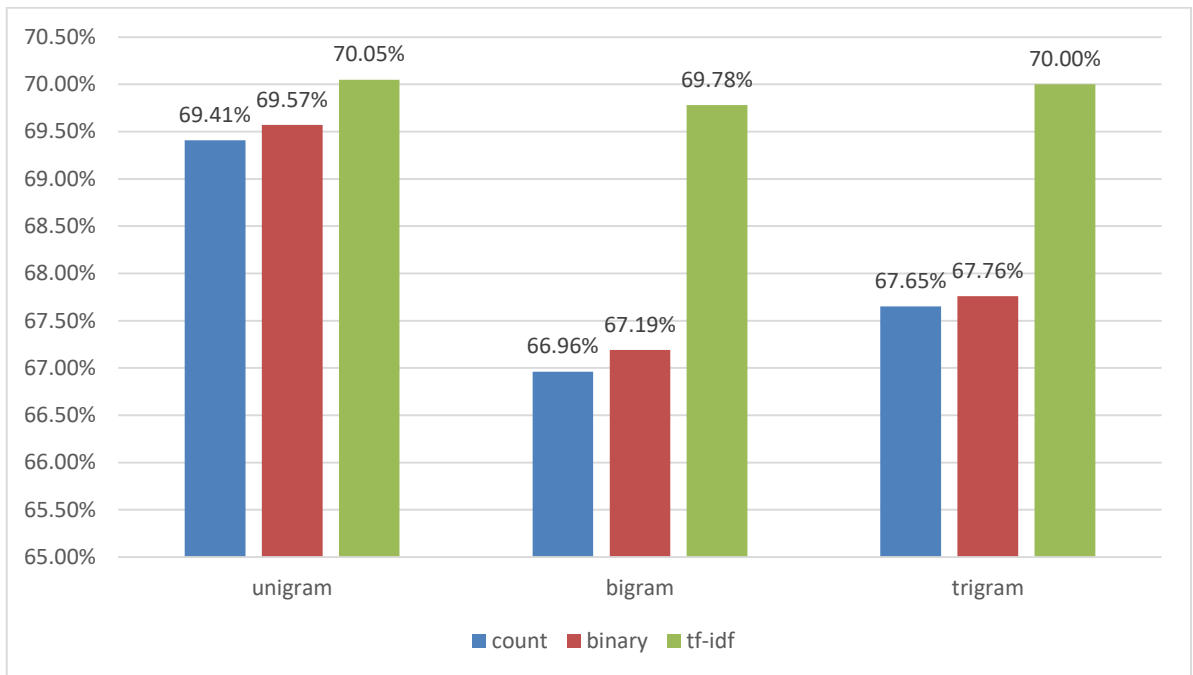
	Count	Binary	Tf-Idf	Trung bình
Unigram	69.41%	69.57%	70.05%	69.68%
Bigram	66.96%	67.19%	69.78%	67.98%
Trigram	67.65%	67.76%	70.00%	68.47%
Trung bình	68.01%	68.17%	69.95%	68.71%

Hàng dọc đầu tiên là danh sách tập từ điển và hàng ngang đầu tiên là danh sách các trọng số tương ứng. Bảng 3.4 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 3.10: Biểu đồ thể hiện kết quả theo trọng số.

Theo hình 3.10 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 3.11: Biểu đồ thể hiện kết quả theo tập từ điển.

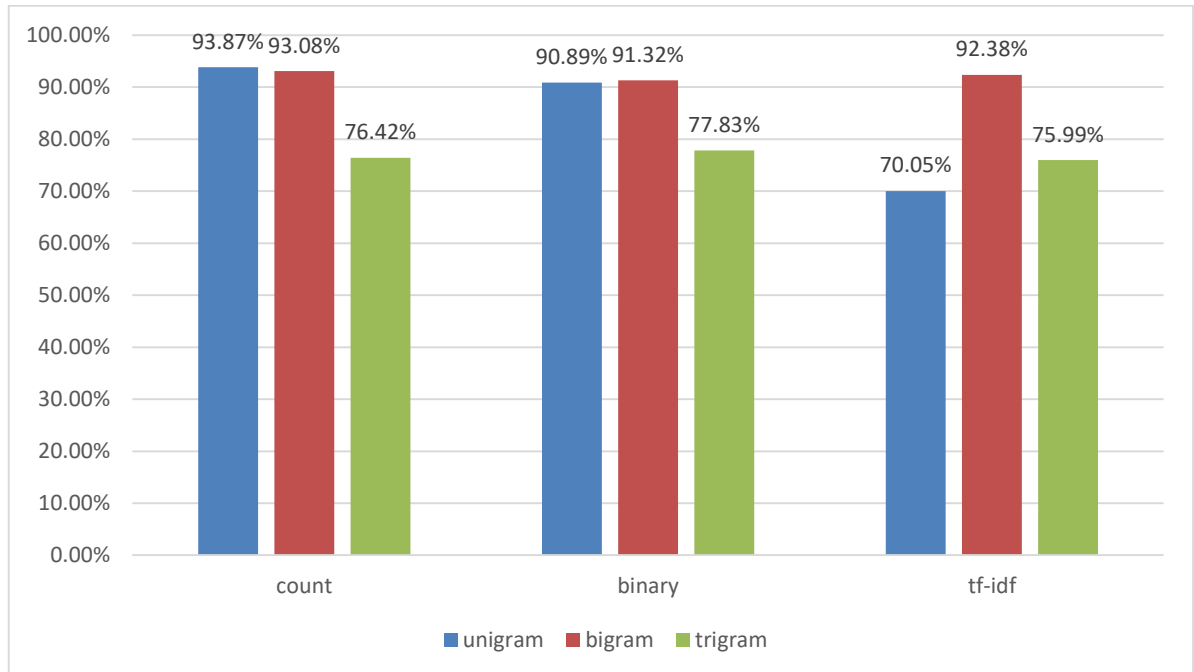
Ngược lại nếu xét trên tập từ điển thì unigram cho kết quả tốt nhất trung bình là 69.68% rồi đến trigram là 68.47% và cuối cùng đến từ điển bigram là 67.98% như biểu đồ hình 3.11.

Nếu 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 từng 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.

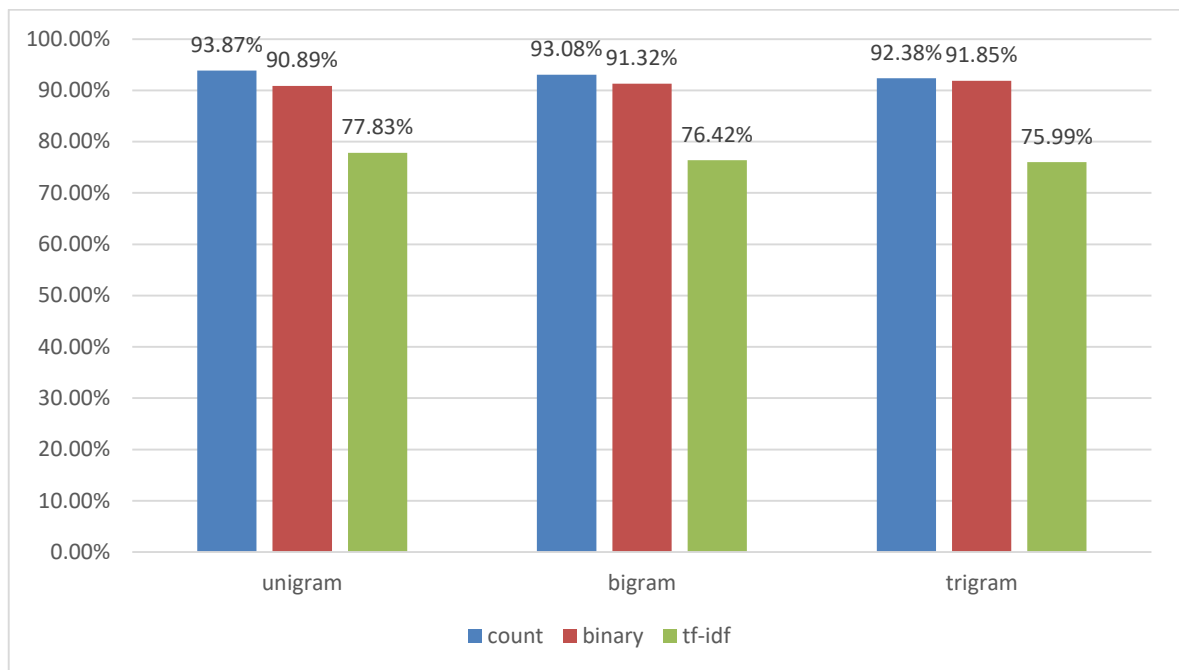
	Count	Binary	Tf-Idf	Trung bình
Unigram	93.87%	90.89%	77.83%	87.53%
Bigram	93.08%	91.32%	76.42%	86.94%
Trigram	92.38%	91.85%	75.99%	86.74%
Trung bình	93.11%	91.35%	76.75%	87.07%

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

Theo hình 3.12 ta thấy nếu xét theo trọng số 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 3.13: 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 đến bigram là 86.94% và cuối cùng đến từ điển bigram là 86.74% như biểu đồ Hình 3.13.

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 mức 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 theo người dùng, việc 1 người có nhiều Status mức độ quan trọng của từ trong 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 việc 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%

Theo Bảng 3.6 độ chính xác cao nhất là 64.45% của tập từ điển trigram với trọng số TF-IDF và thấp nhất là 61.57% là tập từ điển unigram với trọng số lần xuất hiện của từ, độ chênh lệch của hai độ chính xác là 2.88%. Độ chính xác trung bình của cả tập dữ liệu là 62.73%.

Bảng 3.7: 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%

Theo Bảng 3.7 giống như tập dữ liệu 10,000 Status, độ chính xác cao nhất thuộc tập từ điển trigram với trọng số TF-IDF là 67.45% và thấp nhất là 64.77% của hai file với tập từ điển bigram với 2 trọng số lần xuất hiện và Binary, độ chênh lệch của hai độ chính xác là 2.68%. Độ chính xác trung bình của cả tập dữ liệu là 65.99%.

Bảng 3.8: Kết quả độ chính xác của tập dữ liệu với 100,000 Status.

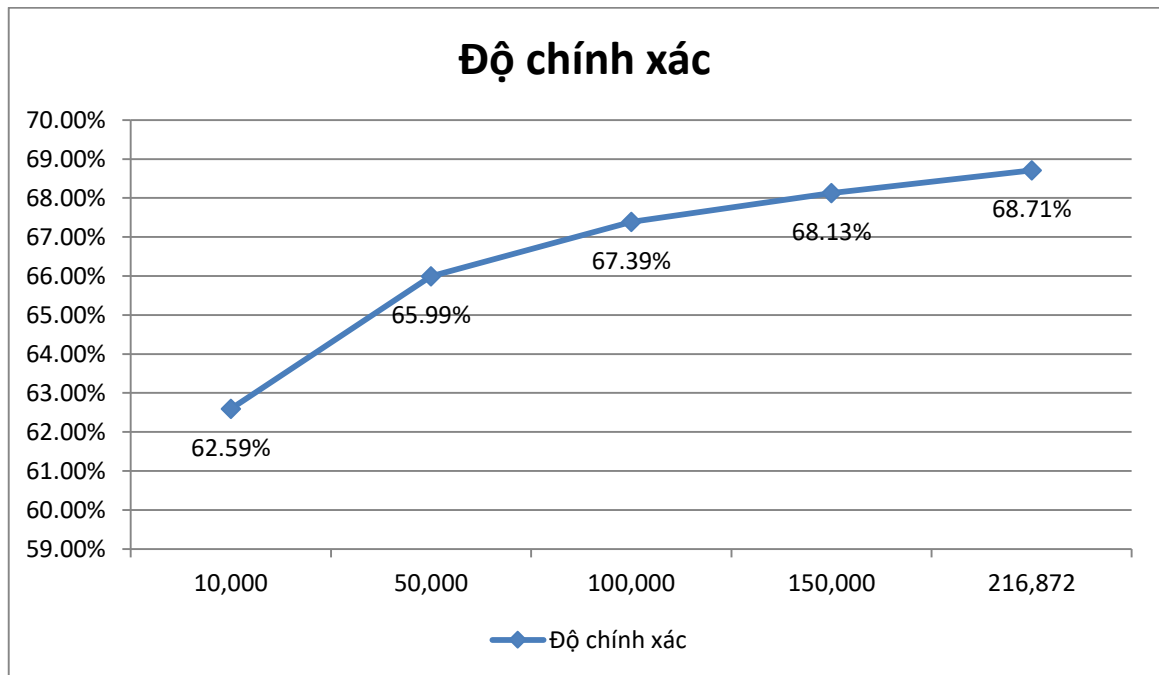
	Count	Binary	Tf-Idf	Trung bình
Unigram	67.68%	67.97%	68.68%	68.11%
Bigram	65.90%	66.10%	68.39%	66.80%
Trigram	66.43%	66.64%	68.72%	67.26%
Trung bình	66.67%	66.90%	68.60%	67.39%

Theo Bảng 3.8, độ chính xác cao nhất vẫn thuộc tập từ điển trigram với trọng số TF-IDF là 68.72% và thấp nhất là 65.90% của tập từ điển bigram với trọng số lần xuất hiện, độ chênh lệch của hai độ chính xác là 2.82%. Độ chính xác trung bình của cả tập dữ liệu là 67.39%.

Bảng 3.9: 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%

Theo Bảng 3.9, độ chính xác cao nhất thuộc tập từ điển trigram với trọng số TF-IDF là 69.58% và thấp nhất là 66.51% của tập từ điển bigram với trọng số lần xuất hiện, độ chênh lệch của hai độ chính xác là 3.07%. Độ chính xác trung bình của cả tập dữ liệu là 68.13%.



Hình 3.14: Biểu đồ kết quả độ chính xác trung bình của từng tập dữ liệu.

Hình 3.14 cho thấy độ chính xác tỉ lệ thuận với số lượng dữ liệu Status. Số lượng càng lớn thì độ chính xác càng cao. Chênh lệch giữa tập dữ liệu lớn và tập dữ liệu 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 các phương pháp thực nghiệm trên tập dữ liệu thu thập được. Các giai đoạn tiền xử lý dữ liệu để xây dựng lên file 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 đoán 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. 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 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.
- ✚ 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 với dữ liệu lớn.
- ✚ Xây dựng hệ thống dự đoán giới tính người dùng mạng xã hội hoàn chỉnh.

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 AAI 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 AAI 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/>