BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

----------o0o----------

# TÌM HIỂU VỀ SUPPORT VECTOR MACHINE CHO BÀI TOÁN PHÂN LỚP QUAN ĐIỂM

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công Nghệ Thông Tin

Sinh viên thực hiện: **Phạm Văn Sơn**

Giáo viên hướng dẫn: **Ths. Nguyễn Thị Xuân Hương**

Mã số sinh viên: **120704**

HẢI PHÒNG 12/2012

# MỤC LỤC

# LỜI CẢM ƠN

Trước hết, em xin chân thành cảm ơn Trường Đại học Dân Lập Hải Phòng. Các Thầy, Cô trong Khoa Công nghệ Thông tin đã tạo điều kiện thuận lợi cho em trong suốt quá trình học tập và làm luận văn tốt nghiệp

Em xin bày tỏ lòng biết ơn sâu sắc của mình đối với Cô Nguyễn Thị Xuân Hương, người đã tận tình hướng dẫn em thực hiện luận văn tốt nghiệp này. Cô đã định hướng cho luận văn, đã giúp sinh viên có một môi trường học thuật để có thể trao đổi ý tưởng, kiến thức đã thu thập được qua đọc sách, tạp chí, tài liệu, qua tìm hiểu các bài giảng, cũng như qua mạng Inernet, đặc biệt Cô đã cho phép sinh viên được tiếp cận với kho tài liệu tương đối đầy đủ, có tính cập nhật cao mà cô đã dày công sưu tầm

Em xin cảm ơn các Thầy, Cô đã quan tâm góp ý và nhận xét quý báu cho bản đồ án của em.

Xin cảm ơn các bạn đã chia sẻ và góp ý cho tôi trong quá trình hoàn thành luận văn

*Hải Phòng, ngày.....tháng.....năm…….*

Sinh viên

***Phạm Văn Sơn***

# MỞ ĐẦU

Trong thời đại hiện nay, sự phát triển như vũ bão của công nghệ thông tin (CNTT) đã kéo theo sự phát triển của nhiều lĩnh vực khác. Có thể nói, CNTT đang làm thay đổi hình hài của nền kinh tế thế giới, giúp nhân loại bước những bước vững chắc đầu tiên trên con đường của kinh tế tri thức, thương mại điện tử.. Ngày nay, con người không còn phải vất vả nhọc nhằn trong công việc thu thập dư liệu vì đã có trợ thủ đắc lực là hệ thống máy tính và mạng truyền số liệu triển khai ở quy mô toàn cầu.

Tuy nhiên, sự phát triển vượt bậc của CNTT đã làm tăng số lượng giao dịch thông tin trên mạng Internet một cách đáng kể, đặc biệt là thư điện tử, tin tức điện tử,... Theo số liệu thống kê từ Brođer et ai (2008) thì cứ sau khoảng 6 đến 10 tháng lượng thông tin đó lại tảng gấp đôi, bên cạnh đó tốc độ thay đổi thông tin cũng cực kỳ nhanh. Hoạt động của các lĩnh vực cũng đặt ra phải xử lý một khối lượng thông tin đồ sộ. Một yêu cầu lớn đặt ra đối với chứng ta là làm sao tổ chức, tìm kiếm thông tin một cách hiệu quả nhất và phân loại thông tin là một trong những giải pháp hợp lý cho yêu cầu này. Nhưng vối một khối lượng thông tin quá lớn và đòi hỏi phải xử lý nhanh thì việc phân loại thủ công là điều không tưởng. Hướng giải quyết là xây dựng các giải pháp cho phép thuật toán hóa và chương trình hóa trên máy tính để có thể tự động phân loại các thông tin trên.

Trong đề tài tốt nghiệp đại học Trường Đại Học Dân Lập Hải Phòng, em thực hiện đề tài ***"TÌM HIỂU VỀ SUPPORT VECTOR MACHINES CHO BÀI TOÁN PHÂN LỚP QUAN ĐIỂM"*** .

***Lý do chọn đề tài***

Vấn đề phân lớp và dự đoán là khâu rất quan trọng trong học máy và trong khai phá dữ liệu, phát hiện trí thức. Kỹ thuật Support Vector Machines (SVM) được đánh giá là công cụ mạnh và tinh vi nhất hiện nay cho những bài toán phân lớp phi tuyến. Nhiều những ứng dụng đã và đang được xây dựng dựa trên kỹ thuật SVM rất hiệu quả.

***Mục đích, đối tượng và phạm vi nghiên cứu***

Trong khuôn khổ luận văn sẽ nghiên cứu phần bài toán phân lớp quan điểm, cơ sở lý thuyết của phương pháp SVM và các vấn đề liên quan. Phân tích những giải pháp cho phép mở rộng và cải tiến để nâng cao hiệu quả ứng dụng của SVM. Đưa kỹ thuật mờ vào SVM cho phép phân chia không gian dữ liệu một cách tốt hơn, nhằm loại bỏ những vùng không được phân lớp bằng SVM thông thường.

Trình bày hướng áp dụng kỹ thuật SVM cũng như những cải tiến, mở rộng của nó vào giải quyết một số các bài toán ứng dụng trong thực tiễn.

Trình bày tổng quan về bài toán phân lớp quan điểm và cụ thể là bài toán phân lớp phân cực để phân chia các tài liệu chứa quan điểm là tích cực hay tiêu cực.

Tìm hiểu dữ liệu quan điểm và viết chương trình thử nghiệm phân lớp phan cực tài liệu sử dụng SVM.

***Ý nghĩa khoa học và thực tiễn***

SVM là một phương pháp phân lớp hiện đại và hiệu quả, nắm chắc phương pháp này sẽ tạo nền tảng giúp chúng ta trong việc phát triển các giải pháp phân loại và dự đoán..., xây dựng được những ứng dụng quan trọng trong thực tế.

Ứng dụng phân lớp SVM cho bài toán phân lớp quan điểm là bài toán đã và đang được nghiên cứu và phát triển rộng rãi và có ý nghĩa cả về học thuật lẫn ứng dụng thực tế.

***Nội dung cơ bản của luận văn bao gồm***

*Chương 2: Tìm hiểu về Support Vector Machine*

*Chương 2:  Bài toán phân lớp quan điểm*

*Chương 3: Chương trình thực nghiệm*

*Phần Kết Luận*

*Phần tài liệu tham khảo*

# CHƯƠNG 1: TÌM HIỂU VỀ SUPPORT VECTOR MACHINE

## 1.1 PHÁT BIỂU BÀI TOÁN

Support Vector Machines (SVM) là kỹ thuật mới đối với việc phân lớp dữ liệu, là phương pháp học sử dụng không gian giả thuyết các hàm tuyến tính trên không gian đặc trưng nhiều chiều, dựa trên lý thuyết tối ưu và lý thuyết thống kê.

Trong kỹ thuật SVM không gian dữ liệu nhập ban đầu sẽ được ánh xạ vào không gian đặc trưng và trong không gian đặc trưng này mặt siêu phẳng phân chia tối ưu sẽ được xác định.

Ta có tập S gồm *e* các mẫu học

$$S = \{(x_1,y_1), (x_2,y_2), (x_3,y_3)\ldots..( x_e,y_e)\} \subseteq (X \text{ x } Y)^e$$

với một vectơ đầu vào *n* chiều $x_i \in R^n$ thuộc lớp I hoặc lớp II (tương ứng nhãn $y_i = 1$ đối với lớp I và $y_i = -1$ đối với lớp II). Một tập mẫu học được gọi là tầm thường nếu tất cả các nhãn là bằng nhau.

Đối với các dữ liệu phân chia tuyển tính, chúng ta có thể xác định được siêu phẳng *f(x)* mà nó có thể chia tập dữ liệu. Khí đó, với mỗi siêu phẳng nhận được ta có: $f(x) \geq 0$ nếu đầu vào *x* thuộc lớp dương, và $f(x) < 0$ nếu *x* thuộc lớp âm

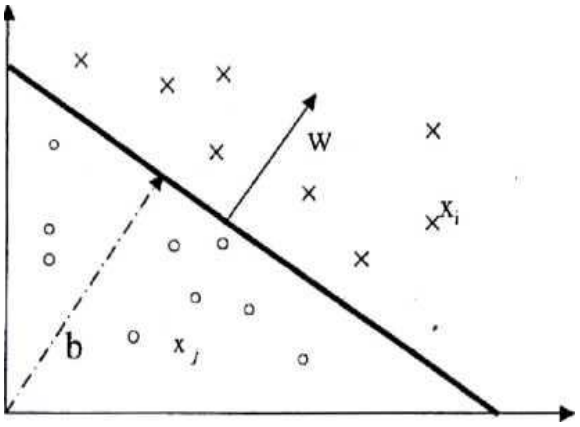
$$f(x) = w.x + b = \sum_{j=1}^{n} w_i x_i + b$$

$$y_i f(x_i) = y_i (w.x_i + b) \geq 0 , \; i=1,....,l$$

trong đó *w* là vector pháp tuyến *n* chiều và *b* là giá trị ngưỡng

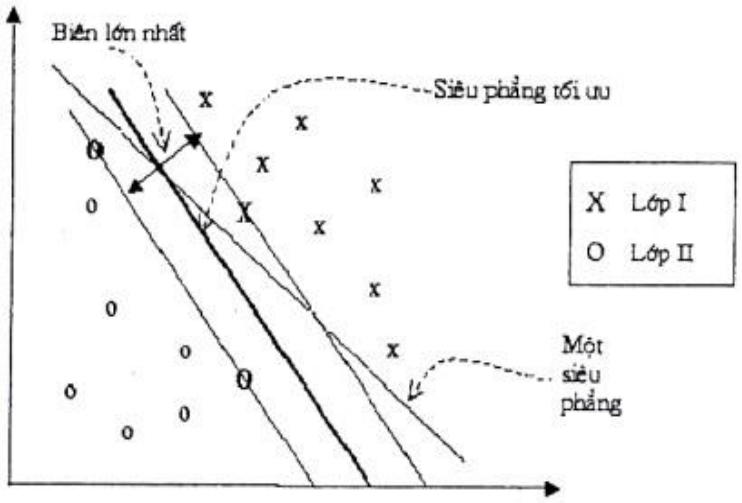
Vector pháp tuyến *w* xác định chiều của siêu phẳng *f(x)*, còn giá trị ngưỡng *b* xác định khoảng cách giữa siêu phẳng và gốc.

.




*Hình 2. 1: Phân tách theo siêu phẳng (w,b) trong không gian 2 chiều của tập mẫu*

Siêu phẳng có khoảng cách với dữ liệu gần nhất là lớn nhất (tức có biên lớn nhất) được gọi là siêu phẳng tối ưu




*Hình 2. 2: Siêu phẳng tối ưu*

Mục đích đặt ra ở đây là tìm được một ngưỡng (*w,b*) phân chia tập mẫu vào các lớp có nhãn 1 (lớp I) và -1 (lớp II) nêu ở trên với khoảng cách là lớn nhất

### 1.1.1 Trình bày tóm tắt về phân lớp dữ liệu

- **Phân lớp dữ liệu** là một kỹ thuật trong khai phá dữ liệu được sử dụng rộng rãi nhất và được nghiên cứu mở rộng hiện nay.
- **Mục đích:** Để dự đoán những nhãn phân lớp cho các bộ dữ liệu hoặc mẫu mới.

*Đầu vào: Một tập các mẫu dữ liệu huấn luyện,với một nhãn phân lớp cho mỗi mẫu dữ liệu*

*Đầu ra: Bộ phân lớp dựa trên tập huấn luyện,hoặc những nhãn phân lớp*

Phân lớp dữ liệu dựa trên tập huấn luyện và các giá trị trong một thuộc tính phân lớp và dùng nó để xác định lớp cho dữ liệu mới

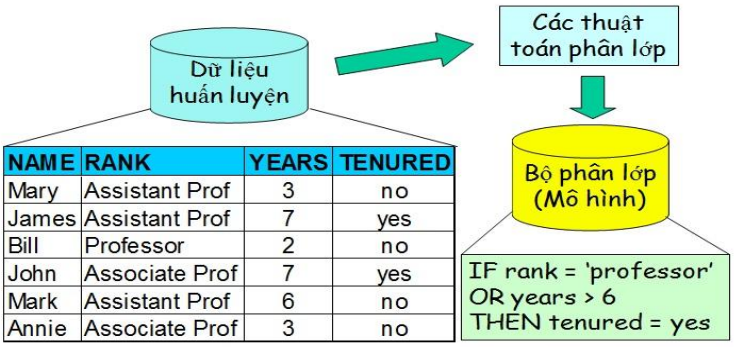
Kỹ thuật phân lớp dữ liệu được tiến hành bao gồm 2 bước:

*Bước 1: Xây dựng mô hình từ tập huấn luyện*

*Bước 2: Sử dụng mô hình – kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới.*

***Bước 1. Xây dựng mô hình***

- Mỗi bộ/mẫu dữ liệu được phân vào một lớp được xác định trước.
- Lớp của một bộ/mẫu dữ liệu được xác định bởi thuộc tính gán nhãn lớp
- Tập các bộ/mẫu dữ liệu huấn luyện - tập huấn luyện - được dùng để xây dựng mô hình.
- Mô hình được biểu diễn bởi các luật phân lớp,các cây quyết định hoặc các công thức toán học.

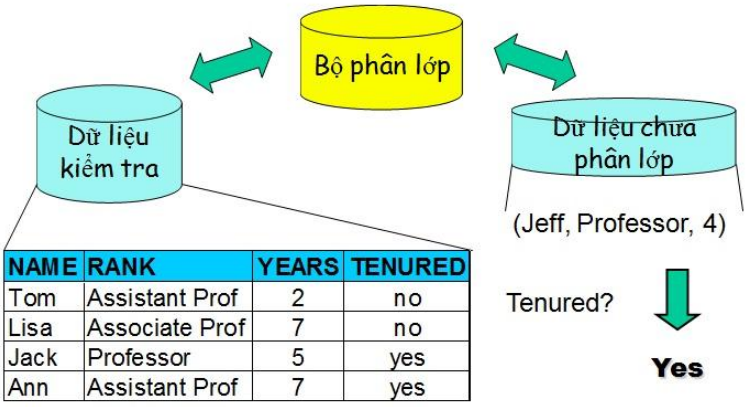



| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Mary | Assistant Prof | 3 | no |
| James | Assistant Prof | 7 | yes |
| Bill | Professor | 2 | no |
| John | Associate Prof | 7 | yes |
| Mark | Assistant Prof | 6 | no |
| Annie | Associate Prof | 3 | no |

*Hình 2.3: Ví dụ xây dựng mô hình*

***Bước 2: Sử dụng mô hình***

- Phân lớp cho những đối tượng mới hoặc chưa được phân lớp
- Đánh giá độ chính xác của mô hình
    - Lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình.
    - Tỉ lệ chính xác bằng phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra




| NAME | RANK | YEARS | TENURED |
|---|---|---|---|
| Tom | Assistant Prof | 2 | no |
| Lisa | Associate Prof | 7 | no |
| Jack | Professor | 5 | yes |
| Ann | Assistant Prof | 7 | yes |

*Hình 2.4: Sử dụng mô hình*

### 1.1.2 Tại sao lại sử dụng thuật toán SVM trong phân lớp dữ liệu

- ✓ SVM rất hiệu quả để giải quyết bài toán dữ liệu có số chiều lớn (ảnh của dữ liệu biểu diễn gene, protein, tế bào)
- ✓ SVM giải quyết vấn đề *overfitting* rất tốt (dữ liệu có nhiễu và tách dời nhóm hoặc dữ liệu huấn luyện quá ít)
- ✓ Là phương pháp phân lớp nhanh
- ✓ Có hiệu suất tổng hợp tốt và hiệu suất tính toán cao

## 1.2 THUẬT TOÁN SVM

### 1.2.1 Giới thiệu

Bài toán phân lớp (*Classification*) và dự đoán (*Prediction*) là hai bài toán cơ bản và có rất nhiều ứng dụng trong tất cả các lĩnh vực như: học máy, nhận dạng, trí tuệ nhân tạo, .v.v . Trong khóa luận này, chúng em sẽ đi sâu nghiên cứu phương pháp Support Vector Machines (SVM), một phương pháp rất hiệu quả hiện nay.

Phương pháp SVM được coi là công cụ mạnh cho những bài toán phân lớp phi tuyến tính được các tác giả Vapnik và Chervonenkis phát triển mạnh mẽ năm 1995. Phương pháp này thực hiện phân lớp dựa trên nguyên lý Cực tiểu hóa Rủi ro có Cấu trúc SRM (*Structural Risk Minimization*), được xem là một trong các phương pháp phân lớp giám sát không tham số tinh vi nhất cho đến nay. Các hàm công cụ đa dạng của SVM cho phép tạo không gian chuyên đổi để xây dựng mặt phẳng phân lớp

### 1.2.2 Định nghĩa

Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là chính xác

Là thuật toán học giám sát (*supervied learning*) được sử dụng cho phân lớp dữ liệu.

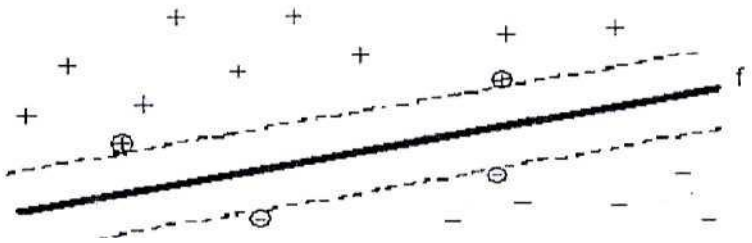
Là 1 phương pháp thử nghiệm, đưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu

SVM là một phương pháp có tính tổng quát cao nên có thể được áp dụng cho nhiều loại bài toán nhận dạng và phân loại

### 1.2.3 Ý tưởng của phương pháp

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. 5: Siêu phẳng phân chia dữ liệu học thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần nhất (điểm được khoanh tròn) là các Support Vector.*

### 1.2.4 Nội dung phương pháp

#### *1.2.4.1 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 $x_i \in R^n$ , 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 $\overrightarrow{x_i}$ trong không gian:

$$\overrightarrow{x_i}.\overrightarrow{w} + b = 0$$

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

Như vậy, f($\vec{X_i}$) biểu diễn sự phân lớp của $\vec{X_i}$ vào hai lớp như đã nêu. Ta nói $y_i$= +1 nếu $\vec{X_i}$ € lớp I và $y_i$ = -1 nếu $\vec{X_i}$ € lớp II . Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau:

Tìm min $\|\vec{w}\|$ với W thỏa mãn điều kiện sau:

$$y_i(\sin(\vec{X_i}.\vec{W} + b)) \geq 1 \text{ với } \forall i \in \overrightarrow{1,n}$$

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 $1/\|\vec{w}\|$. 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 đùng để tối ưu hóa kết quả.

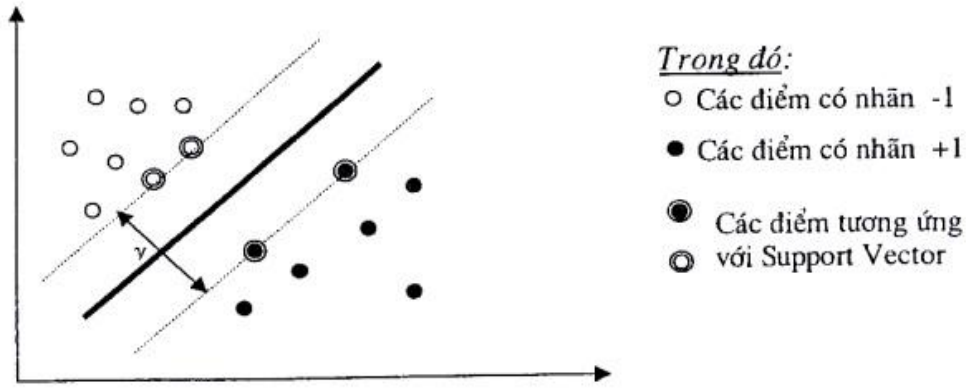
***TÓM LẠI:*** trong trường hợp nhị phân phân tách tuyến tính, việc phân lớp được thực hiện qua hàm quyết định *f(x) = sign(<w.x> + b)*, hàm này thu được bằng việc thay đổi vectơ chuẩn *w*, đây là vectơ để cực đại hóa viền chức năng

Việc mở rộng SVM để phân đa lớp hiện nay vẫn đang được đầu tư nghiên cứu. Có một phương pháp tiếp cận để giải quyết vấn để này là xây dựng và kết hợp nhiều bộ phân lớp nhị phân SVM (Chẳng hạn: trong quá trình luyện với SVM, bài toán phân m lớp có thể được biến đổi thành bài toán phân 2*m lớp, khi đó trong mỗi hai lớp, hàm quyết định sẽ được xác định cho khả năng tổng quát hóa tối đa). Trong phương pháp này có thể đề cập tới hai cách là *một-đổi-một, một-đối-tất cả*

### *1.2.4.2 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

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




Hình 2. 6: Minh họa bài toán 2 phân lớp bằng phương pháp SVM

***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

### *1.2.4.3 Bài toán nhiều phân lớp với SVM*

Để phân nhiều lớp thì kỹ thuật SVM nguyên thủy sẽ chia không gian dữ liệu thành 2 phần và quá trình này lặp lại nhiều lần. Khi đó hàm quyết định phân dữ liệu vào lớp thứ i của tập n , 2-Iớp sẽ là:

$$f_i(x) = w^i_i x + b_i$$

Những phần tử *x* là support vector sẽ thỏa điều kiện

$$f_i(x) = \begin{cases} +1 \text{ nếu thuộc lớp i} \\ \\ -1 \text{ nếu thuộc phần còn lại} \end{cases}$$

Như vậy, bài toán phân nhiều lớp sử dụng phương pháp SVM hoàn toàn có thể thực hiện giống như bài toán hai lớp. Bằng cách sử dụng chiến lược *"một- đối-một"*(*one - against - one*).

Giả sử bài toán cần phân loại có k lớp (k > 2), chiến lược *"một-đối-một"*sẽ tiến hành k(k-l)/2 lần phân lớp nhị phân sử dụng phương pháp SVM. Mỗi lớp sẽ tiến hành phân tách với k-1 lớp còn lại để xác định k-1 hàm phân tách dựa vào bài toán phân hai lớp bằng phương pháp SVM.

#### *1.2.4.4 Các bước chính của phương pháp SVM*

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

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

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

Thực hiện việc kiểm tra chéo để xác định các thám 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.

Kiểm thử tập dữ liệu Test

# CHƯƠNG 2: BÀI TOÁN PHÂN LỚP QUAN ĐIỂM

## 2.1 SỰ KIỆN (Facts) VÀ QUAN ĐIỂM (Opinions)

Thông tin dạng văn bản có thể chia thành 2 loại chính:

- Sự kiện là những biểu hiện khách quan về các thực thể, các sự kiện và các thuộc tính của họ

  VD: "Tôi vừa mua chiếc ô tô màu xanh này"

- Quan điểm là những biểu hiện chủ quan mô tả tình cảm, đánh giá hay cảm xúc của con người đối với các thực thể, sự kiện và thuộc tính của chúng: thể hiện dạng tích cực, tiêu cực hay trung lập.

  VD: "Chiếc ô tô này thật đẹp, và dễ sử dụng"

## 2.2 NHU CẦU VỀ THÔNG TIN QUAN ĐIỂM VÀ NHẬN XÉT

"Những gì người khác nghĩ" đã luôn luôn là một phần quan trọng trong việc cung cấp thông tin cho quá trình ra quyết định của hầu hết chúng ta. Trước khi Internet trở lên phổ biến, chúng ta thường yêu cầu bạn bè hay người thân giới thiệu một thợ cơ khí tự động hoặc yêu cầu tài liệu tham khảo thư liên quan đến xin việc từ các đồng nghiệp, hoặc tư vấn tiêu dùng. Ngày nay, Internet và Web đã giúp cho chúng ta có thể dễ dàng tiếp cận các ý kiến và kinh nghiệm của những người khác mà không nhất thiết phải là những người quen biết cá nhân, không phải là các nhà phê bình chuyên nghiệp nổi tiếng, những người mà chúng ta chưa bao giờ nghe nói tới trong không gian rộng lớn. Và ngược lại, ngày càng nhiều và nhiều hơn nữa những người sẵn sàng cung cấp các ý kiến của mình cho những người khác qua Internet.

Theo như hai cuộc khảo sát của hơn 2000 người Mỹ trưởng thành mỗi: 81% người dùng Internet (hoặc 60% người Mỹ) đã thực hiện nghiên cứu trực tuyến về một sản phẩm ít nhất một lần 20% (15% của tất cả các người Mỹ) làm như vậy trong một ngày. Trong số các độc giả đánh giá trực tuyến của nhà hàng, khách sạn, và các dịch vụ khác nhau (ví dụ như, các cơ quan du lịch hoặc bác sĩ), giữa 73% và 87% báo cáo đánh giá đã có một ảnh hưởng đáng kể mua hàng của họ. Người tiêu dùng sẵn sàng trả từ 20% đến 99% một mục được đánh giá 5 sao cao hơn so với

một mục đánh giá 4 sao. 32% đã cung cấp một đánh giá về một sản phẩm, dịch vụ thông qua một hệ thống xếp hạng trực tuyến, trong đó có 18% của công dân trực tuyến cao cấp, có đăng một bình luận trực tuyến hoặc xem xét về một sản phẩm hay dịch vụ.

Thống kê nhanh chỉ ra rằng việc tiêu thụ hàng hóa và dịch vụ không phải là động cơ duy nhất khi người dùng tìm kiếm hoặc thể hiện ý kiến trực tuyến. Sự cần thiết của những thông tin chính trị cũng là một yếu tố quan trọng. Ví dụ, trong một cuộc khảo sát hơn 2500 người Mỹ trưởng thành, Rainie và Horrigan nghiên cứu có 31% người Mỹ - trên 60 triệu người - 2006 người dùng Internet vận động tranh cử, là những người thu thập thông tin về cuộc bầu cử năm 2006 trực tuyến và trao đổi nhận xét thông qua email.

Trong số này:

- 28% nói rằng nguyên nhân chính cho các hoạt động trực tuyến này để thu nhận được quan điểm từ bên trong cộng đồng của họ, và 34% cho biết một lý do chính là để nhận được quan điểm từ bên ngoài cộng đồng của họ.
- 27% đã xem đánh giá trực tuyến cho sự tán thành hoặc xếp hạng của các tổ chức bên ngoài;
- 28% cho biết rằng hầu hết các trang web mà họ sử dụng để chia sẻ quan điểm, nhưng 29% nói rằng phần lớn các trang web mà họ sử dụng thách thức quan điểm của họ, chỉ ra rằng nhiều người không chỉ đơn giản là tìm kiếm để xác nhận các quan điểm có trước của họ
- 8% đăng bình luận trực tuyến bình luận chính trị riêng của họ.

Đối với người dùng tìm kiếm sự tin cậy trong những lời khuyên và tư vấn trực tuyến quan tâm đến việc xây dựng một hệ thống mới để xử lý trực tiếp các quan điểm trước tiên là phân loại chúng. Theo Horrigan thống kê rằng trong khi đa số người sử dụng internet của Mỹ cho rằng kinh nghiệm tích cực trong nghiên cứu sản phẩm trực tuyến, 58% cho rằng thông tin trực tuyến là thiếu, khó tìm, khó hiểu và hoặc quá nhiều. Vì vậy, nhu cầu có một hệ thống để hỗ trợ người tiêu dùng tìm kiếm thông tin là rất cần thiết.

Các nhà cung cấp sản phẩm ngày càng chú ý hơn đến sự quan tâm mà người dùng cá nhân thể hiện trong các ý kiến trực tuyến về sản phẩm và dịch vụ, và sự ảnh hưởng như xu thế sử dụng.

Với sự bùng nổ của nền tảng Web 2.0 như các blog, diễn đàn thảo luận, *peer-to-peer* mạng, và các loại khác nhau của các mạng xã hội... Một lượng đông đảo người dùng gia tăng chưa từng có và có quyền chia sẻ kinh nghiệm và ý kiến của riêng họ về bất kỳ sản phẩm hoặc dịch vụ, là tích cực hay tiêu cực. Khi các công ty lớn đang ngày càng nhận ra, những tiếng nói của người tiêu dùng có thể vận dụng rất lớn ảnh hưởng trong việc hình thành ý kiến của người tiêu dùng khác, cuối cùng, để trung thành với thương hiệu của họ, họ quyết định  mua,và vận động cho chính thương hiệu của họ... Công ty có thể đáp ứng với những hiểu biết của người tiêu dùng mà họ tạo ra thông qua điều khiển phương tiện truyền thông xã hội và phân tích các thông điệp maketing của họ, định vị thương hiệu, phát triển sản phẩm và các hoạt động phù hợp khác.

Tuy nhiên, các nhà phân tích ngành công nghiệp lưu ý rằng việc tận dụng các phương tiện truyền thông mới cho mục đích theo dõi hình ảnh sản phẩm đòi hỏi cần phải có công nghệ mới.

Các nhà tiếp thị luôn luôn cần giám sát các phương tiện truyền thông cho thông tin liên quan đến thương hiệu của mình, cho dù đó là đối với các hoạt động quan hệ công chúng, vi phạm gian lận, hoặc tình báo cạnh tranh. Nhưng phân mảnh các phương tiện truyền thông và thay đổi hành vi của người tiêu dùng đã loại trừ các phương pháp giám sát truyền thống. Technorati ước tính rằng 75.000 blog mới được tạo ra mỗi ngày, cùng với 1,2 triệu bài viết mỗi ngày, trong đó có nhiều ý kiến người tiêu dùng thảo luận về sản phẩm và dịch vụ.

Vì vậy, không chỉ có cá nhân, mà các công ty, các tổ chức  đều quan tâm đến một hệ thống có khả năng tự động phân tích quan điểm của người tiêu dùng

## 2.3 MÁY TÌM KIẾM QUAN ĐIỂM / NHẬN XÉT

Tạo hệ thống có thể xử lý thông tin chủ quan một cách hiệu quả đòi hỏi phải khắc phục một số thách thức mới. Để một ứng dụng sẽ điền vào thông tin quan trọng và phổ biến cần thiết, có hạn chế sự chú ý vào blog tìm kiếm hoặc xem xét các loại tổng quát hơn của tìm kiếm đã được mô tả ở trên.

Sự phát triển của các ứng dụng tìm kiếm quan điểm hoặc ý kiến hoàn chỉnh có thể liên quan đến việc tấn công lẫn nhau trong những vấn đề sau đây.

- Khi ứng dụng được tích hợp vào một máy tìm kiếm, sau đó cần phải xác định xem thông tin chủ quan mà người dùng thực tế tìm kiếm. Điều này có thể không là một vấn đề khó khăn: trong truy vấn của loại này có thể gồm các thuật ngữ như: "*review*" "*reviews*" hoặc "*opinion*" , hoặc ứng dụng có thể cung cấp một hộp kiểm tra để người dùng để họ có thể chọn trực tiếp các đánh giá cần tìm. Tuy nhiên, truy vấn phân loại là một bài toán khó và là một vấn đề thách thức
- Bên cạnh đó vấn đề vẫn còn mở là xác định những tài liệu tại chỗ liên quan đến một truy vấn định hướng quan điểm. Đây cũng là một một thách thức trong việc thiết lập quyết định là đồng thời hoặc sau đó, là các tài liệu hoặc các phần của tài liệu chứa tài liệu chứa nhận xét hoặc quan điểm.
- Khi đã có các tài liệu mục tiêu, người ta vẫn còn phải đối mặt với vấn đề xác định quan điểm tổng thể thể hiện trong đó và/hoặc ý kiến cụ thể liên quan các đặc trưng riêng biệt hoặc các khía cạnh của các đối tượng hoặc các chủ đề trong câu hỏi, khi cần thiết.
- Cuối cùng, hệ thống cần phải trình bày thông tin quan điểm thu được trong một số trang tóm tắt hợp lý.

## 2.4 LỊCH SỬ CỦA PHÂN TÍCH QUAN ĐIỂM VÀ KHAI THÁC QUAN ĐIỂM

Lĩnh vực phân tích quan điểm (*sentiment analysis*) hay khai thác quan điểm (*opinion mining*) gần đây đã thu hút được sự quan tâm rộng rãi của các nhà nghiên cứu. Năm 2001 bắt đầu đánh dấu sự lan rộng nhận thức về các vấn đề nghiên cứu và cơ hội nâng cao phân tích tình cảm và khai thác quan điểm.

***Các nhân tố được nghiên cứu gồm:***

- ✓ Sự gia tăng của các phương pháp học máy, xử lý ngôn ngữ tự nhiên và khôi phục thông tin.
- ✓ Sự sẵn có của các tập dữ liệu đào tạo cho các thuật toán học máy, sự phát triển của Internet, cụ thể là sự phát triển của tập hợp các trang Web thu thập các ý kiến và quan điểm.

Thực hiện những thách thức trí tuệ, thương mại và các ứng dụng thông minh trong lĩnh vực này.

Thuật ngữ khai thác quan điểm (Dave et al. 2003) là các công cụ khai thác quan điểm sẽ xử lý một tập hợp các kết quả tìm kiếm cho một đối tượng nhất định, sinh ra một danh sách các thuộc tính sản phẩm (chất lượng, đặc trưng, vv) và các quan điểm tổng hợp về chúng (kém, bình thường, tốt).

"*Phân tích quan điểm*"là cụm từ song song của "*khai thác quan điểm*"ở những khía cạnh nhất định (Das và Chen Tong, 2001). *"Phân tích quan điểm"*và *"khai thác quan điểm"*biểu thị cùng một lĩnh vực nghiên cứu

## 2.5 XU HƯỚNG NGHIÊN CỨU GẦN ĐÂY

Gần đây, khai thác quan điểm đã trở thành chủ đề nóng giữa các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin. Có khá nhiều các bài báo được xuất bản và những ứng dụng khác nhau có sử dụng hệ thống đánh giá quan điểm được phát triển và đưa vào trong hoạt động thương mại. Các tiếp cận chủ yếu với bài toán này là:

- ✓ *Phân lớp quan điểm thông qua việc xác định từ, cụm từ chỉ quan điểm*
- ✓ *Xác định quan điểm với các thể hiện trong từng thuộc tính của đối tượng cần tìm kiếm quan điểm.*

### 2.5.1 Xác định cụm từ, quan điểm

Những từ, cụm từ chỉ quan điểm là những từ ngữ được sử dụng để diễn tả cảm xúc, ý kiến người viết, những quan điểm chủ quan đó dựa trên những vấn đề mà anh ta hay cô ta đang tranh luận. Việc rút ra những từ, cụm từ chỉ quan điểm là giai đoạn đầu tiên trong hệ thống đánh giá quan điểm, vì những từ, cụm từ này là những chìa khóa cho công việc nhận biết và phân loại tài liệu sau đó.

Ứng dụng dựa trên hệ thống đánh giá quan điểm hiện nay tập trung vào các từ chỉ nội dung câu: danh từ, động từ, tính từ và phó từ. Phần lớn công việc sử dụng từ loại để rút chúng ra (Hu và Liu, 2004 , Turney, 2002). Việc gán nhãn từ loại cũng được sử dụng trong công việc này, điều này có thể giúp cho việc nhận biết xu hướng quan điểm trong giai đoạn tiếp theo. Những kĩ thuật phân tích ngôn ngữ tự nhiên khác như xóa: *stopwords, stemming* cũng được sử dụng trong giai đoạn tiền xử lý để rút ra từ, cụm từ chỉ quan điểm

### 2.5.2 Sử dụng tính từ và phó từ

Những hệ thống hiện tại dùng để nhận biết những từ chỉ quan điểm hay xu hướng quan điểm tập trung chủ yếu vào các tính từ và phó từ vì chúng được xem là sự biểu lộ rõ ràng nhất của tính chủ quan ( Hatzivassiloglou and McKeown, 1997, Wiebe and Bruce, 1999 ).

Hu và Liu (2004) áp dụng việc gán nhãn từ loại và kĩ thuật xử lý ngôn ngữ tự nhiên nhằm rút ra những tính từ cũng như những từ chỉ quan điểm. Phương pháp của họ dựa vào việc phân loại dựa trên dấu hiệu quan điểm về sản phẩm:

- Định nghĩa một câu mà chứa một hay nhiều dấu hiệu sản phẩm và từ chỉ quan điểm được xem là một câu chỉ quan điểm.
- Với mỗi câu trong dữ liệu chỉ quan điểm, rút ra tất cả những tính từ được coi là những từ chỉ quan điểm.
- Kết quả thực nghiệm việc rút ra những câu đánh giá quan điểm có độ chính xác (*precision*) khoảng 64.2% và *recall* là 69.3%.
- Sử dụng WordNet (Fellbaum, 1998) để xác định các tính từ được rút ra mang chiều hướng tích cực (*positive*) hay tiêu cực (*negative*).

Trong WordNet, các tính từ được tổ chức thành các cụm từ lưỡng cực, nửa cụm thứ hai phần đầu là từ trái nghĩa của cụm thứ nhất. Mỗi nửa cụm là phần đầu của tập từ đồng nghĩa chính, tiếp theo là tập từ đồng nghĩa kèm theo, đại diện cho ngữ nghĩa tương tự như những tính từ quan trọng. Ngược với cách tiếp cận dựa trên từ điển, họ sử dụng định hướng quan điểm của những từ đồng nghĩa và từ trái nghĩa để dự đoán định hướng của các tính từ. Họ bắt đầu với một danh sách khởi đầu gồm 30 tính từ thông dụng được chọn thủ công (bằng tay). Sau đó sử dụng WordNet để

dự đoán định hướng của tất cả các tính từ trong danh sách từ quan điểm được rút ra bằng cách tìm kiếm qua cụm lưỡng cực để tìm ra liệu các từ đồng nghĩa hay trái nghĩa có trong danh sách khởi đầu hay không. Khi định hướng của tính từ được dự đoán, nó sẽ được bổ sung vào danh sách khởi đầu và có thể được sử dụng để xác định định hướng của các tính từ khác. Trong phương pháp này, danh sách khởi đầu sẽ dần tăng lên khi sự định hướng của các tính từ được nhận dạng, và khi nó ngừng gia tăng, tức qui mô của danh sách khởi đầu trùng với qui mô của danh sách từ chỉ quan điểm, thì tất cả định hướng của các tính từ đã được nhận biết và quá trình này kết thúc.

Những từ quan điểm thường tập trung chủ yếu vào hai từ loại: tính từ và phó từ  vì vậy càng nhận dạng chính xác được nhiều hai loại từ này hệ thống càng có độ chính xác cao

### 2.5.3 Sử dụng các động từ

Các tính từ và phó từ đóng một vai trò quan trọng trong việc phân tích quan điểm và là các loại từ có lợi thế trong việc nhận biết định hướng và rút ra các từ chỉ quan điểm trong các nghiên cứu hiện nay. Tuy nhiên, các loại từ khác, ví dụ như động từ cũng được sử dụng để diễn tả cảm xúc hay ý kiến trong các bài viết.

Nasukawa và Yi (2003) xem xét rằng bên cạnh các tính từ và phó từ, thì các động từ cũng có thể diễn tả quan điểm trong hệ thống đánh giá quan điểm của họ. Họ phân loại các động từ có liên quan đến quan điểm thành 2 loại. Loại thứ nhất trực tiếp thể hiện quan điểm tích cực hay tiêu cực, theo lý giải của họ thì “beat” trong “X beats Y” . Loại thứ hai không thể hiện quan điểm trực tiếp nhưng dẫn đến những quan điểm , giống như “is” trong “X is good” .

Họ sử dụng gán nhãn từ loại dựa trên mô hình Markov (HMM) (Manning and Schutze, 1999) và phân tích cú pháp nông dựa trên luật (Neff et al., 2003) cho bước tiền xử lý. Sau đó họ phân tích tính phụ thuộc về mặt cú pháp giữa các cụm từ và tìm kiếm các cụm từ có một từ chỉ quan điểm mà nó bổ nghĩa hoặc được bổ nghĩa bởi một thuật ngữ chủ thể

### 2.5.4 Xác định chiều hướng, cụm từ, quan điểm

Trong phân tích quan điểm, xu hướng của những từ, cụm từ trực tiếp thể hiện quan điểm, cảm xúc của người viết bài. Phương pháp chính để nhận biết xu hướng quan điểm của những từ, cụm từ chỉ cảm nghĩ là dựa trên thống kê hoặc dựa trên từ vựng

## 2.6 NHIỆM VỤ CỦA PHÂN TÍCH QUAN ĐIỂM

Phân tích quan điểm là những nghiên cứu nhằm phát hiện ra quan điểm hay xu hướng của người dùng dựa trên các kỹ thuật liên quan đến vấn đề xử lý ngôn ngữ tự nhiên. Có hai hướng tiếp cận chính cho bài toán này là:Phân lớp quan điểm (*Sentiment Classification*) và trích quan điểm (*Sentiment Extraction*)

Trích quan điểm: bao gồm 3 nhiệm vụ chính là:

- *Trích các đặc trưng đối tượng có nhận xét trong mỗi quan điểm.*
- *Xác định có hay không các quan điểm trong các đặc trưng là positive, negative hay neutral (phụ thuộc vào định dạng của các quan điểm)*
- *Nhóm các cụm từ cùng nghĩa đặc trưng*

## 2.7 BÀI TOÁN PHÂN LỚP QUAN ĐIỂM

Phân lớp là quá trình "nhóm" các đối tượng "giống" nhau vào "một lớp" dựa trên các đặc trưng dữ liệu của chúng. Tuy nhiên, phân lớp là một hoạt động tiềm ẩn trong tư duy con người khi nhận dạng thế giới thực, đóng vai trò quan trọng làm cơ sở đưa ra các dự báo, các quyết định. Phân lớp và cách mô tả các lớp giúp cho tri thức được định dạng và lưu trữ trong đó

Khi nghiên cứu một đối tượng, hiện tượng, chúng ta chỉ có thể dựa vào một số hữu hạn các đặc trưng của chúng. Nói cách khác, ta chỉ xem xét biểu diễn của đối tượng, hiện tượng trong một không gian hữu hạn chiều, mỗi chiều ứng với một đặc trưng được lựa chọn. Khi đó, phân lớp dữ liệu trở thành phân hoạch tập dữ liệu thành các tập con theo một tiêu chuẩn nhận dạng được.

Nhiệm vụ phân lớp quan điểm được xem xét với hai tiếp cận chính là:

- Phân lớp câu chứa quan điểm
- Phân lớp tài liệu chứa quan điểm.

Phân lớp câu/tài liệu chứa quan điểm có thể được phát biểu như sau: Cho một câu hay một tài liệu chứa quan điểm, hãy phân loại xem câu hay tài liệu đó thể hiện quan điểm mang xu hướng tích cực(*positive*) hay tiêu cực (*negative*), hoặc trung lập (*neutral*).

Theo Bo Pang và Lillian Lee (2002) phân lớp câu/tài liệu chỉ quan điểm không có sự nhận biết của mỗi từ/ cụm từ chỉ quan điểm. Họ sử dụng học máy có giám sát để phân loại những nhận xét về phim ảnh. Không cần phải phân lớp các từ hay cụm từ chỉ quan điểm, họ rút ra những đặc điểm khác nhau của các quan điểm và sử dụng thuật toán Naïve Bayes (NB), Maximum Entropy (ME) và Support Vector Machine (SVM) để phân lớp quan điểm. Phương pháp này đạt độ chính xác từ 78, 7% đến 82, 9%.

***Input:*** Cho một tập các văn bản chứa các ý kiến đánh giá về một đối tượng nào đó.

***Output:*** Mỗi văn bản được chia vào một lớp theo mức độ phân cực (*polarity*) về tiếp cận ngữ nghĩa nào đó (tích cực, tiêu cực hay trung lập).

Phân lớp tài liệu theo hướng quan điểm thật sự là vấn đề thách thức và khó khăn trong lĩnh vự xử lý ngôn ngữ đó chính là bản chất phức tạp của ngôn ngữ của con người, đặc biệt là sự đa nghĩa và nhập nhằng nghĩa của ngôn ngữ. Sự nhập nhằng này rõ ràng sẽ ảnh hưởng đến độ chính xác bộ phân lớp của chúng ta một mức độ nhất định. Một khía cạnh thách thức của vấn đề này dường như là phân biệt nó với việc phân loại chủ đề theo truyền thống đó là trong khi những chủ đề này được nhận dạng bởi những từ khóa đứng một mình, quan điểm có thể diễn tả một cách tinh tế hơn. Ví dụ câu sau: "*Làm thế nào để ai đó có thể ngồi xem hết bộ phim này ?*"không chứa ý có nghĩa duy nhất mà rõ ràng là nghĩa tiêu cực. Theo đó, quan điểm dường như đòi hỏi sự hiểu biết nhiều hơn, tinh tế hơn

### 2.7.1 Phân cực quan điểm và mức độ phân cực

- Mức độ phân cực: *positive*/*negative*/*neutral*
- Nhận xét về sản phẩm, dịch vụ: Like/ dislike/ So so
- Nhận xét về phim ảnh thumbs up/ thumbs down
- Nhận xét về quan điểm chính trị:like to win/ unlike to win Liberal/conservative
- Phân loại bài báo là good new/ bad new.

Các bài toán liên quan đến phân lớp phân cực quan điểm:

- Xác định sự phân cực của văn bản (tài liệu/câu) chứa quan điểm: tích cực, tiêu cực hay trung tính.

  VD: Thông qua nhận xét: "This laptop is great".

- Xác định một đoạn thông tin "khách quan" là tốt hoặc xấu =>thách thức liên quan đến phân tích quan điểm.

  VD: "The stock prise rose"

- Phân biệt giữa câu "chủ quan"và "khách quan"

Rating inference (*ordinal regression*): Sắp xếp các quan điểm theo nhiều mức:

- Sắp xếp các đánh giá từ theo nhiều mức: VD: 1 sao đến 5 sao. Hay theo mức độ phân cực: rất thích, thích, bình thường, không thích,…
- Khi phân loại vào 3 lớp: *positive*, *negative*, *neutral*: neutral được coi là giá trị trung bình giữa *positive* và *negative*.
- Nhãn "*neutral*": một số được sử dụng như là lớp khách quan(thiếu quan điểm).
- Theo Cabral và Hortacsu, 2006: nhãn *neutral* có thể gần *negative* hơn vì con người có xu hướng phản ứng mạnh với nhận xét *negative*: 40% so với nhận xét neutral là 10%.

### 2.7.2 Nhiệm vụ của bài toán phân lớp quan điểm

Bài toán phân lớp quan điểm được biết đến như là bài toán phân lớp tài liệu với mục tiêu là phân loại các tài liệu theo định hướng quan điểm.

Đã có rất nhiều tiếp cận khác nhau được nghiên cứu để giải quyết cho loại bài toán này. Để thực hiện, về cơ bản có thể chia thành hai nhiệm vụ chính như sau:

- *Trích các đặc trưng nhằm khai thác các thông tin chỉ quan điểm để phục vụ mục đích phân loại tài liệu theo định hướng ngữ nghĩa.*
- *Xây dựng mô hình để phân lớp các tài liệu.*

### 2.7.3 Xây dựng mô hình phân lớp để phân loại tài liệu

Trong phân tích quan điểm, xu hướng của những từ, cụm từ trực tiếp thể hiện quan điểm, cảm xúc của người viết bài. Phương pháp chính để nhận biết xu hướng quan điểm của những từ, cụm từ chỉ cảm nghĩ là dựa trên thống kê hoặc dựa trên từ vựng. Với nhiệm vụ phân lớp các tài liệu, đã có rất nhiều các phương pháp học máy thống kê được sử dụng cho mục đích này, như là: Naive Bayes, phân loại Maximum Entropy, học máy giám sát SVM, cây quyết định,…

*Thuật toán gồm 4 bước:*

***Bước 1:*** *Xác định các n-gram, các đặc trưng được lọc qua toàn bộ tập dữ liệu.*

***Bước 2:*** *Tính toán tần số xuất hiện của các n-gram tích cực, tiêu cực và tính trọng số của các n-gram.*

***Bước 3****: Chọn n-gram thỏa mãn ngưỡng và có trọng số cao cũng như loại bỏ các bigram không có ý nghĩa cho việc phân loại.*

***Bước 4:*** *Tính toán độ chính xác của quá trình huấn luyện của bộ phân lớp*

# CHƯƠNG III: CHƯƠNG TRÌNH THỰC NGHIỆM

## 3.1 DỮ LIỆU THỬ NGHIỆM

Trong đồ án này, chúng tôi sử dụng dữ liệu từ những bài viết về đánh giá bộ phim gồm 700 nhận xét tích cực và 700 nhận xét tiêu cực. Dữ liệu này được cung cấp bởi tác giả Lillian Lee (http://www.cs.cornell.edu/people/pabo/movie-review-data/) Các dữ liệu này đã dược loại bỏ các chỉ số đánh giá và rút ra thông tin trong nguyên văn từ các tài liệu gốc định dạng HTML, xử lý các dấu chấm câu như các mục của từ vựng riêng biệt.

Chúng tôi tập trung vào những đặc trưng dựa trên những từ đơn **<unigram>** và cặp 2 từ **<bigram>.**

## 3.2 CÔNG CỤ SỬ DỤNG

### 3.2.1 Công cụ sinh SRIML

SRILM là bộ công cụ để xây dựng và áp dụng các mô hình ngôn ngữ thống kê , chủ yếu là để sử dụng trong nhận dạng tiếng nói, gắn thẻ thống kê và phân khúc, và dịch máy thống kê. Bộ công cụ này được phát triển bởi "Phòng thí nghiệm và nghiên cứu công nghệ giọng nói SRI" từ năm 1995, có thể chạy trên nền tảng Linux cũng như Windows.

SRILM bao gồm các thành phần sau:

- Một tập hợp các thư viện C++ giúp cài đặt mô hình ngôn ngữ, hỗ trợ cấu trúc dữ liệu và các chức năng tiện ích nhỏ.
- Một tập hợp các chương trình thực thi thực hiện nhiệm vụ xây dựng mô hình ngôn ngữ, đào tạo và thử nghiệm mô hình ngôn ngữ trên dữ liệu, gắn thẻ hoặc phân chia văn bản, ...

Bộ công cụ SRILM có rất nhiều chương trình con, để xây dựng mô hình ngôn ngữ ta sử dụng chương trình Ngram

Chương trình Ngram thống kê tần số xuất hiện của các cụm Ngram. Kết quả của việc thống kê được ghi lại vào một tệp hoặc sử dụng chúng để xây dựng mô hình ngôn ngữ. Kết quả của việc thống kê được ghi lại theo định dạng sau:

**ngram –count –ordern -interpolate -text <dataFile> -lm <outputFile>**

Trong đó:

- ***order n***: thiết lập độ dài lớn nhất của các cụm Ngram sẽ thống kê bằng n. Giá trị mặc định nếu không thiết lập tham số này là n = 3
- ***interpolaten:*** với n nhận các giá trị là 1, 2, 3, 4, 5, 6, 7, 8, hoặc 9. Tính toán tần số của các cụm Ngram có độ dài là n bằng cách nội suy từ các cụm Ngram có độ dài nhỏ hơn.
- ***text<dataFile>:*** File dữ liệu cần thống kê tần số các cụm Ngram. Tệp văn bản này có thể chứa mỗi câu trên một dòng. Kí hiệu kết thúc và bắt đầu dòng mới sẽ được tự động thêm vào nếu trong tệp đầu vào chưa có. Các dòng trống trong tệp này cũng bị loại bỏ.
- ***lm<outputFile>:*** xây dựng mô hình ngôn ngữ truy hồi từ các tần số vừa thống kê, sau đó ghi lại vào tệp fileketqua theo định dạng ở trên.

### 3.2.2 Ngôn ngữ lập trình java

Java là một ngôn ngữ lập trình dạng lập trình hướng đối tượng (OOP). Khác với phần lớn ngôn ngữ lập trình thông thường, thay vì biên dịch mã nguồn thành mã máy hoặc thông dịch mã nguồn khi chạy, Java được thiết kế để biên dịch mã nguồn thành bytecode, bytecode sau đó sẽ được môi trường thực thi (runtime environment) chạy. Bằng cách này, Java thường chạy nhanh hơn những ngôn ngữ lập trình thông dịch khác như Python, Perl, PHP,…

Cú pháp Java được vay mượn nhiều từ C & C++ nhưng có cú pháp hướng đối tượng đơn giản hơn và ít tính năng xử lý cấp thấp hơn.

Một số đặc điểm nổi bật của java

- Máy ảo java
- Thông dịch
- Độc lập nền
- Hướng đối tượng

Đa nhiệm, đa luồng

### 3.2.3 Công cụ phân lớp dữ liệu SVMLight

Bộ công cụ phân lớp SVM-light viết trên C được phát triển bởi Joachims Thorste với các đặc điểm chính sau:

Các tính năng chính của chương trình

- Tối ưu hóa thuật toán nhanh
- Giải quyết nhanh các vấn đề phân loại và hồi quy đối với các kết quả đầu ra đa biến
- Hỗ trợ các phương pháp nhận dạng mẫu....

SVM-light bao gồm các thành phần chính:

- SVMTlearn: huấn luyện mô hình
- SVMTagger: gán nhãn phân lớp
- SVMTeval: đánh giá kết quả.
- SVMClassicfy: kiểm thử kết quả

Thực hiện:

Huấn luyện mô hình:

**svm-learn [-option] train_file model_file**

- Trong đó:

  ***train_file***: bao gồm dữ liệu huấn luyện.

  - Tên file của ***train_file*** có thể là bất kỳ.
  - Phần mở rộng của phải do người dùng tự đặt nhưng phải giới hạn 3 ký tự.

  ***model_file***: chứa model được xây dựng dựa trên dữ liệu huấn luyện của SVM.

- Với phân lớp văn bản , dữ liệu huấn luyện là các tài liệu thu thập
- Mỗi dòng là thể hiện của một tài liệu.
- Mỗi đặc trưng thể hiện một thuật ngữ (từ) trong một tài liệu
  - Nhãn và mỗi đặc trưng được cách nhau bởi dấu cách.
  - Đặc trưng: cặp giá trị của đặc trưng được sắp xếp theo giá trị số tăng dần.
- Giá trị đặc trưng sử dụng: VD:tf-idf (trọng số của tần xuất từ).

## 3.3 Kết quả thực nghiệm

Các bước thực hiện

***Bước 1:*** sử dụng công cụ N-gram để sinh ra các file dữ liệu chứa các N-gram của tài liệu chứa quan điểm. Ở đây, chúng tôi sử dụng uni-gram (1-gram) và Bi-gram (2-gram).

***Bước 2*:** Từ tập dữ liệu này, trước khi được sử dụng để huấn luyện và kiểm thử cần qua một số bước lọc bỏ các đặc trưng không tốt.

Bước thứ nhất, lọc bỏ các từ vô nghĩa (stop word), và các ký tự đặc biệt như {! , ? , / , @ . # , “, ....}

Bước tiếp theo là lọc bỏ các đặc trưng theo tần số. Những đặc trưng có tần số xuất hiện trong dữ liệu huấn luyện thấp hơn một giá trị nào đó (đối với unigram sẽ là nhỏ hơn 3 và bigram là nhỏ hơn 7) sẽ bị loại bỏ. Bước cuối cùng được thực hiện sau khi đã gán các trọng số cho từng đặc trưng.

***Bước 3*: Gán nhãn cho mỗi N-gram trong tập dữ liệu huấn luyện để lấy thông tin phân loại: các nhận xét chứa quan điểm tích cực được gán nhãn 1, các nhận xét chứa quan điểm tiêu cực được gán nhãn -1.**

Chương trình sau khi gán nhãn cho các câu, phân loại các tập dữ liệu huấn luyện và đánh giá sẽ tiến hành chọn các đặc trưng là các từ cho đầu vào của thuật toán SVM

Để thực hiện phân lớp tài liệu quan điểm, chúng tôi chia tập dữ liệu thành hai tập con là tập huấn luyện (train)  và tập kiểm thử (test)

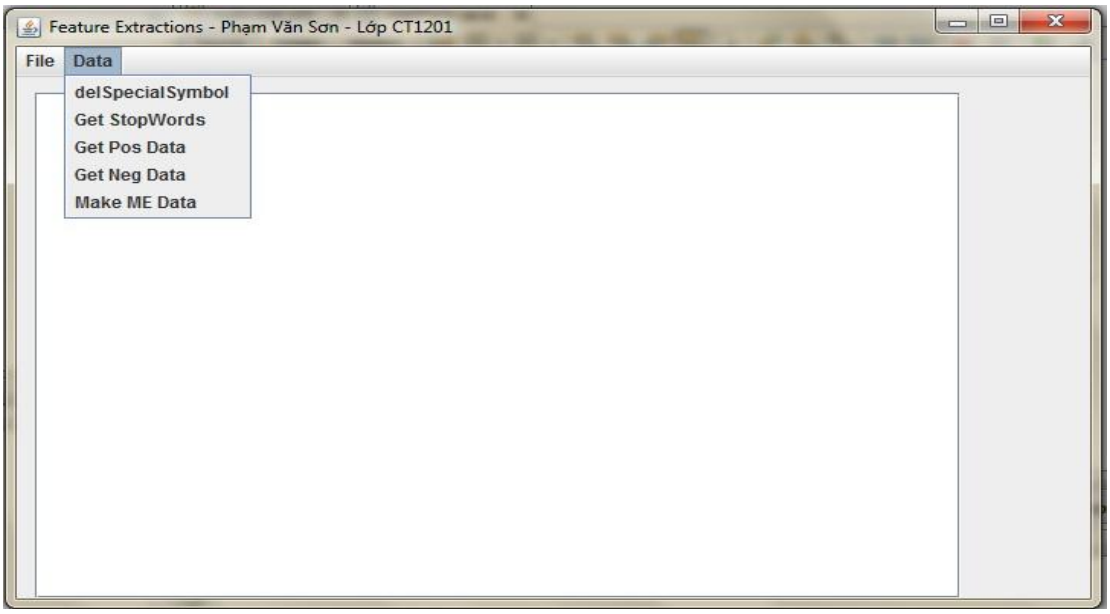
Tập huấn luyện gồm có 550 nhận xét tích cực và 550 nhận xét tiêu cực.

Tập kiểm thử (test) gồm có 150 nhận xét tích cực và 150 nhận xét tiêu cực.

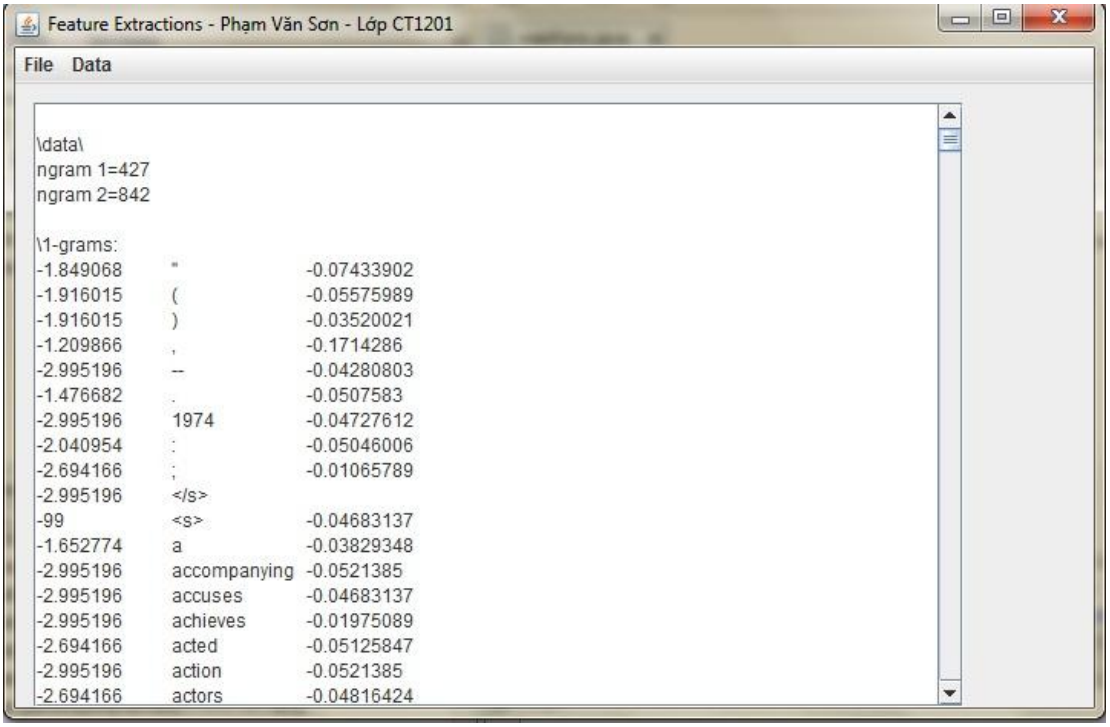
Kết quả thực hiện phân lớp Support Vector Machine với các đặc trưng Uni-gram và Bigram như sau:

| **Đặc trưng** | **Uni-gram** | **Bi-gram** |
|---|---|---|
| Độ chính xác (Precision) | 91,38 % | 56,49% |
| Độ phản hồi (Recall) | 91,54% | 58% |

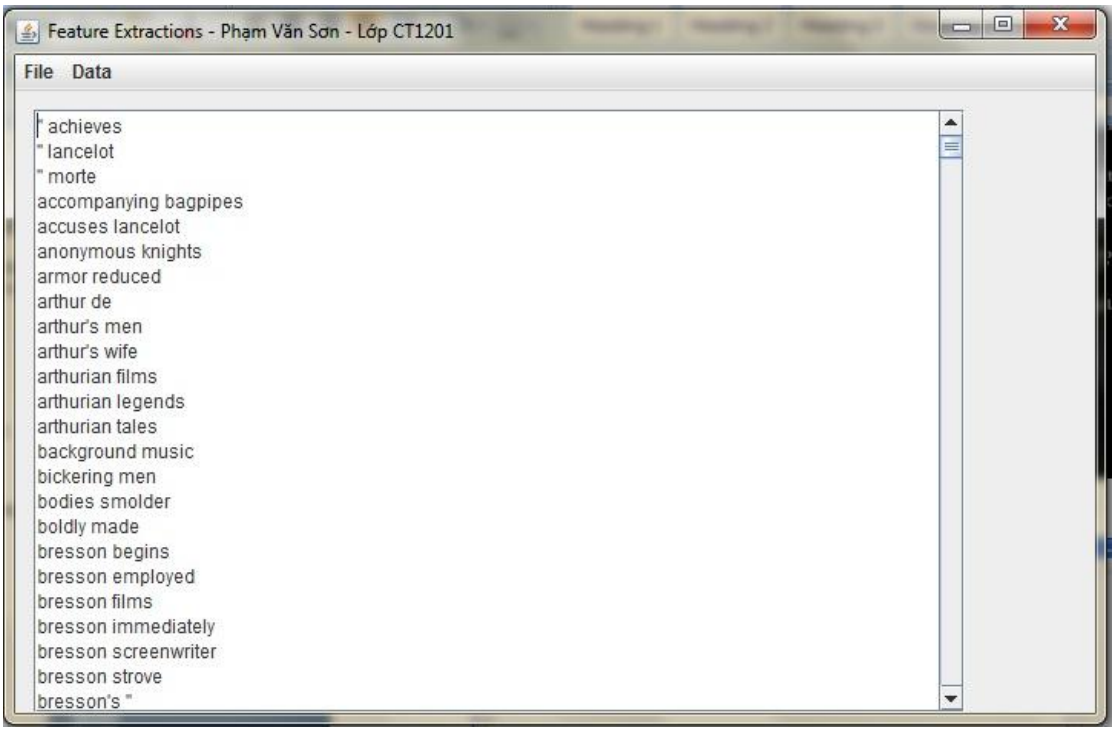
Chương trình trích đặc trưng n-gram và tạo dữ liệu cho phân lớp SVM để phân lớp các bình luận là tích cực hay tiêu cực.



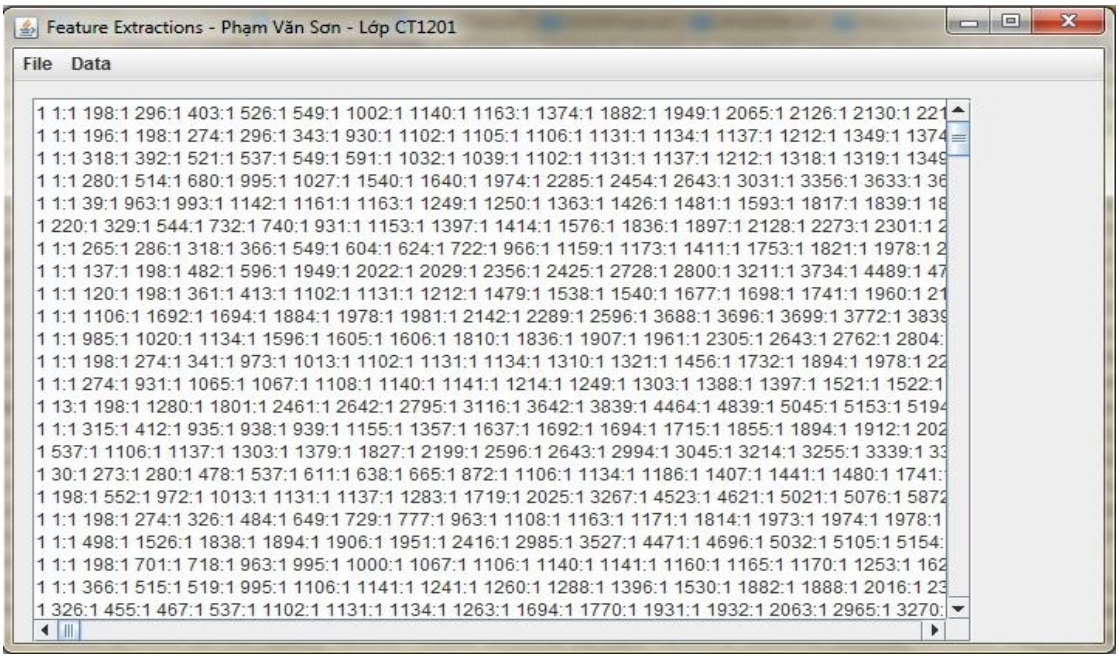

3. 1: Giao diện chính của chương trình



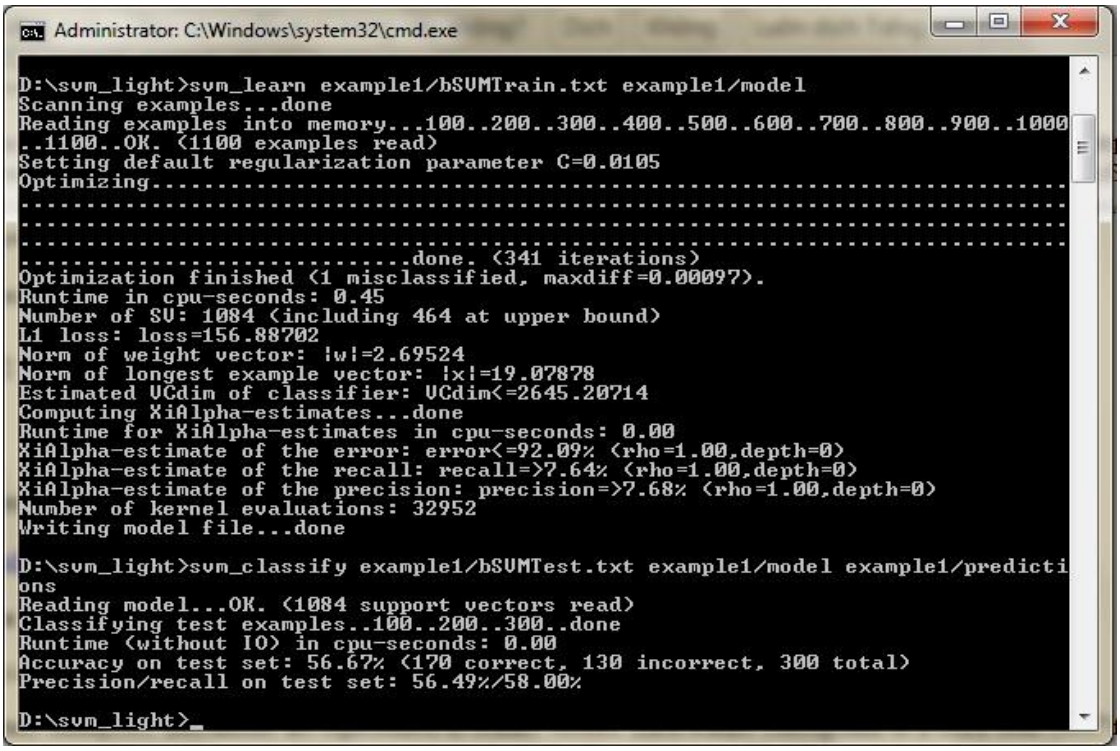

3.2: Mở file dữ liệu đầu vào




3.3: Hiển thị dữ liệu dùng để chạy Get Pos Data




3.4: Dữ liệu cho phân lớp SVM




3.5: Hình ảnh khi chạy Lệnh SVM trong môi trường DOS

# KẾT LUẬN

Luận văn hướng tới mục tiêu phân lớp dữ liệu đạt độ chính xác cao, tuy đã xem xét được tất cả các mục tiêu như trong phần giới thiệu nhưng do thời gian có hạn, nên một số vấn đề vẫn chưa hoàn chỉnh. Tuy nhiên, luận văn cũng đạt được một số kết quả: .

Nghiên cứu và trình bày cơ sở của lý thuyết của phương pháp học máy.

Trình bày phương pháp SVM. Đây là một phương pháp phân lớp hiệu quả được nghiên cứu nhiều nhất trong thời gian qua.

Phân tích những giải pháp cho phép mở rộng và cải tiến để nâng cao hiệu quả ứng dụng của SVM:

- Cài đặt được một số công cụ giúp đỡ cho việc xây dựng mô hình ngôn ngữ như: chuẩn hóa văn bản, tách từ bằng ngôn ngữ Java.
- Cài đặt được chương trình để trích đặc trưng và tạo dữ liệu cho phân lớp SVM.
- Tìm kiếm và sử dụng bộ dữ liệu phân lớp tài liệu chứa quan điểm.
- Cài đặt và chạy thành công bộ mã nguồn mở Srilm trên môi trường Linux
- Sử dụng bộ công cụ mã nguồn mở SRILM để xây dựng mô hình ngôn ngữ cho dữ liệu đầu vào.

Do thời gian có hạn, nên hiện tại luận văn mới chỉ nghiên cứu được trích đặc trưng n-gram từ các bình luận và sử dụng phân lớp SVM để phân lớp các bình luận là tích cực hay tiêu cực. Trong thời gian tới, tôi sẽ tiếp tục nghiên cứu trích các đặc trưng khác cho bài toán này và các phương pháp phân lớp thống kê khác.

# TÀI LIỆU THAM KHẢO


1. **Ths. Nguyễn Thị Xuân Hương và Ths. Lê Thụy về "phân tích quan điểm và một số hướng tiếp cận" . Hội nghị khoa học lần thứ nhất, 2012, trường ĐHDL Hải Phòng**
2. **Nghiên cứu thuật toán phân lớp nhị phân và ứng dụng cho bào toán Protein Folding – Nguyễn Quang Phước – Trường Đại học Khoa học tự nhiên TPHCM**
3. **Bo Pang and Lillian Lee và Shivakumar Vaithyanathan. Thumbs up Sentiment Classification using Machine Learning Techniques.**
4. **http://en.wikipedia.org/wiki/Support_vector_machine**
5. **http://www.cs.cornell.edu**
6. **http://svmlight.joachims.org/**
7. **ftp://ftp.cs.cornell.edu/pub/smart/english.stop**
8. **http://www.speech.sri.com/projects/srilm/download.html**