Bộ giáo dục và đào tạo
Trường đại học dân lập hải phòng
-------o0o-------

# TÌM HIỂU VỀ MAXIMUM ENTROPY 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 Thị Hà
Giáo viên hướng dẫn:Ths Nguyễn Thị Xuân Hương
Mã sinh viên: 120797

Hải Phòng - 2012

# MỤC LỤC

# LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới Thạc sĩ Nguyễn Thị Xuân Hương (Trường Đại học Dân lập Hải Phòng) đã chỉ bảo và hướng dẫn tận tình cho em trong suốt quá trình tìm hiểu và thực hiện khóa luận này.

Em xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới các thầy, cô đã dạy dỗ em trong suốt quá trình học tập tại trường Đại học Dân lập Hải Phòng cũng như những giúp đỡ, những động viên trong suốt quá trình làm khóa luận.

Và con xin gửi lời cảm ơn và biết ơn vô hạn tới bố, mẹ, những người thân yêu của đã nuôi nấng, dạy dỗ và luôn là chỗ dựa tinh thần cho con trong cuộc sống cũng như trong học tập.

Mặc dù em đã cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Em kính mong nhận được sự cảm thông và tận tình chỉ bảo, góp ý của quý Thầy Cô và các bạn.

*Em xin chân thành cảm ơn!*

Hải Phòng, ngày…..tháng…..năm…….

Sinh viên

***Phạm Thị Hà***

# MỞ ĐẦU

Ngày nay, với sự phát triển mạnh mẽ của Internet, các hình thức kết nối và chia sẻ thông tin trong cộng đồng mạng ngày càng phát triển đã thu hút một lượng lớn người dùng tham gia. Thông qua đó họ trao đổi, chia sẻ thông tin, thảo luận các vấn đề và sở thích cùng quan tâm. Một số mạng xã hội phổ biến trên thế giới như: Facebook, Twitter và ở Việt Nam như: Zing, Go.vn có số lượng người tham gia ngày càng đông đảo. Một số các hình thức thể hiện khác cũng rất phát triển như các phản hồi trên các diễn đàn, các trang dịch vụ và các trang tin tức,. . .

Các thông tin được chia sẻ và thảo luận thông qua mạng xã hội thuộc rất nhiều chủ đề trong các lĩnh vực kinh tế, chính trị, xã hội. Từ đó hình thành nên các luồng thông tin chủ đạo (thể hiện như xu hướng của quan điểm) của cộng đồng đối với việc đánh giá một vấn đề, hay một sản phẩm, dịch vụ nào đó. Các quan điểm, xu hướng này sẽ có tác động mạnh mẽ đến định hướng, quan điểm của người dùng khác.

Việc nghiên cứu các phương pháp học máy cho bài toán phân lớp quan điểm đã và đang thu hút được một lượng lớn các nghiên cứu của các nhà khoa học trong lĩnh vực xử lý ngôn ngữ tự nhiên và khai phá dữ liệu. Các kết quả của nó được nghiên cứu trong lĩnh vực này đã có nhiều ứng dụng trên thực tế như: các hệ thống lấy ý kiến phản hồi khách hàng, các nhận xét, đánh giá được tích hợp trong các hệ thống phản hồi trực tuyến,… Chính vì lý do đó, em đã chọn đề tài "**Tìm hiểu về Maximum Entropy cho bài toán phân lớp quan điểm**" cho khóa luận tốt nghiệp của mình.

Nội dung của khóa luận được tổ chức thành ba chương như sau:

**Chương 1:** Trình bày bài toán phân lớp quan điểm, nhiệm vụ của bài toán phân lớp quan điểm.

**Chương 2:** Trình bày về mô hình và thuật toán Entropy cực đại cho bài toán phân lớp quan điểm.

**Chương 3:** Trình bày những kết quả đánh giá thử nghiệm của khóa luận áp dụng cho bài toán phân lớp quan điểm.

Cuối cùng là phần kết luận.

# Chương 1: BÀI TOÁN PHÂN LỚP QUAN ĐIỂM

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

Những thông tin nhận xét 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 lấy những thông tin này từ bạn bè, người thân hay các chuyên gia tư vấn tiêu dùng về sản phẩm dịch vụ quan tâm. Với sự bùng nổ của 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 Khảo sát Kelsey group và Horrigan, 2008:*

- 81% người dùng Internet (60% người Mỹ) nghiên cứu trực tuyến về một sản phẩm ít nhất một lần.
- 20% (15% của tất cả các người Mỹ) thực hiện trong 1 ngày.
- 73% và 87% người dùng bị ảnh hưởng đáng kể đến việc mua hàng.
- Người tiêu dùng sẵn sàng trả từ 20% đến 99% một mục được đánh giá 5 sao ( 32 % với mục 4 sao).
- 18% của công dân trực tuyến cao cấp có một bình luận trực tuyến về sản phẩm hay dịch vụ.

Với sự bùng nổ các dịch vụ web: blog, diễn đàn thảo luận, peer-to-peer mạng, và các loại khác nhau của các mạng xã hội…

*Thống kê của FaceBook:*

- Có hơn 500 triệu người dùng ở trạng thái hoạt động (active), mỗi người có trung bình 130 bạn (friends), trao đổi qua lại trên 900 triệu đối tượng .

*Twitter (5/2011):* có hơn 200 triệu người dùng

- Một ngày có: hơn 300 nghìn tài khoản mới, trung bình hơn 190 triệu tin nhắn, xử lý trung bình khoảng 1.6 tỷ câu hỏi.

*Theo ước tính của Technorati mỗi ngày có:*

- 75.000 blog mới được tạo ra.
- 1, 2 triệu bài viết.

*Ở Việt Nam* các mạng xã hội: zing.vn, go.vn… thu hút được động đảo người dùng tham gia.

Tại đây 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.

## 1.2 BÀI TOÁN PHÂN LỚP QUAN ĐIỂM

Bài toán 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) theo định hướng ngữ nghĩa (tích cực, tiêu cực hay trung lập).

Phân lớp tài liệu theo định 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.

## 1.3 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 phục vụ mục đích phân loại tài liệu theo định hướng ngữ nghĩa.
- Xây dựng mô hình để phân lớp các tài liệu.

*1.3.1. Trích các đặc trưng*

Trích 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. 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.

*1.3.2 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: Naïve Bayes, phân loại maximum Entropy, máy vector tựa SVM, cây quyết định…

Thuật toán gồm các bước sau:

***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 n-gram không có ý nghĩa cho việc phân loại.
- Bước 4: Tính toán độ chính xác của quá trình huấn luyện của bộ phân lớp.

Trong đó:

– Xét trong một văn bản, Ngram là một cụm từ gồm nhiều từ liên tiếp cùng xuất hiện trong văn bản đó. Và do đó ***Bigram*** là một cụm từ gồm 2 từ liên tiếp cùng xuất hiện trong văn bản đó. Nếu độ dài tính theo từ của một văn bản là L thì số N-gram được sinh ra là :

$$\frac{L * (L - N + 1)}{2}$$

Như vậy, N càng nhỏ thì số lượng N-gram sinh ra càng lớn.

– ***Xây dựng các đặc trưng*****:** sử dụng mô hình ngôn ngữ N-gram để xây dựng các mệnh đề thông tin ngữ cảnh, từ đó xây dựng các đặc trưng trước khi đưa vào huấn luyện mô hình.

Có được tập hợp các N-gram, ta tiến hành xây dựng các mệnh đề thông tin ngữ cảnh. Mệnh đề mô tả thông tin ngữ cảnh là một mệnh đề chỉ ra văn bản hiện tại chứa một N-gram nào đó. Ví dụ,

[document has *i_love_it]*

Theo cách mà nguyên lý Entropy đã cung cấp để xây dựng đặc trưng: một đặc trưng là sự kết hợp giữa mệnh đề mô tả thông tin ngữ cảnh

và nhãn của lớp tương ứng với văn bản. Ví dụ, với *y* là lớp "*int*" (interesting) ta có một đặc trưng như sau:

$$f_{document\_has\_i\_love\_it,\,\text{int}}(x,y)=\begin{cases}1 & \text{If y=int and document\_has\_i\_love\_it}\\ 0\end{cases}$$

Đặc trưng này sẽ có giá trị đúng nếu văn bản hiện tại chứa cụm từ "*I love it*" và nó được gán nhãn "*int*".

Cần chú ý rằng, số lượng các mệnh đề thông tin ngữ cảnh sinh ra nhỏ hơn số lượng các N-gram(vì có những N-gram trùng nhau cũng xuất hiện trong một văn bản) và cũng không bằng số lượng các đặc trưng.

– ***Lựa chọn đặc trưng***

*i) Chiến lược loại bỏ stop-word*

Bản chất của các ngôn ngữ tự nhiên là luôn có các câu, từ xuất hiện nhiều nhưng không mang nhiều ý nghĩa để phân loại. Trong tiếng Anh gọi đó là ***stop-word***. Stop-word không những dư thừa, khi kết hợp với các từ khác để xây dựng đặc trưng chúng còn gây ra hiện tượng overfitting. Qua thử nghiệm trên một bộ phân lớp văn bản trên tiếng Anh, sau khi lọc stop-word độ chính xác huấn luyện(trainning accuracy) tăng lên đáng kể. Vì vậy loại bỏ stop-word là rất cần thiết.

Vậy làm thế nào để loại bỏ stop-word. Khóa luận này nghiên cứu một phương pháp khá hiệu quả, đó là sau khi sinh N-gram, và loại bỏ theo quy tắc:

- Loại bỏ các n-gram là stop-word: điều này có lợi với các ngôn ngữ mà đơn vị nhỏ nhất không phải là từ hơn là câu. Ví dụ, câu "he is in Hanoi" sinh ra 2-gram "is_in" chứa 2 stop-word là "is" và "in".

*ii) Đặt ngưỡng*

Thực tế cho thấy, có những mệnh đề thông tin ngữ cảnh xuất hiện nhiều lần trong một văn bản và những mệnh đề thông tin ngữ cảnh xuất hiện rất ít lần. Ví dụ trong câu *"Oil Prices are Escalating":*

[document has *Oil Prices are*]

Để loại bỏ những mệnh đề thông tin ngữ cảnh không có nhiều ý nghĩa này, chiến lược lọc đặt ngưỡng chỉ đơn giản đặt ngưỡng cho sự xuất hiện của một mệnh đề thông tin ngữ cảnh trong toàn bộ tập mệnh đề thông tin ngữ cảnh: nếu số lần xuất hiện nằm ngoài một khoảng nào đó thì bị loại bỏ.

– ***Huấn luyện mô hình***

Sau khi đã xây dựng được tập các đặc trưng ta tiến hành huấn luyện mô hình. Ở bước này chính là lúc áp dụng các thuật toán ước lượng tham số để tìm ra tập trọng số λ(mỗi một đặc trưng $f_i$ sẽ được gán một trọng số $\lambda_i$). Một văn bản mới có một tập các đặc trưng, chính tập trọng số sẽ quyết định mức độ quan trọng của các đặc trưng và chúng ảnh hưởng trực tiếp đến quá trình phân lớp cho văn bản mới, từ đó dự đoán một cách chính xác lớp cho văn bản đó.

Trong quá trình huấn luyện mô hình, chúng ta cần tiến hành đánh giá độ chính xác của bộ phân lớp. Nói cách khác, trong quá trình này cần đánh giá khả năng đoán nhận của mô hình thông qua *độ chính xác*(accuracy) của quá trình huấn luyện.

# Chương 2: MÔ HÌNH ENTROPY CỰC ĐẠI

## 2.1 GIỚI THIỆU

Mô hình Entropy cực đại là mô hình dựa trên xác suất có điều kiện cho phép tích hợp các thuộc tính đa dạng từ dữ liệu mẫu nhằm hỗ trợ quá trình phân lớp.

Tư tưởng chủ đạo của nguyên lý Entropy cực đại rất đơn giản: ta phải xác định môt phân phối mô hình sao cho phân phối đó tuân theo mọi giả thiết đã quan sát từ thực nghiệm, ngoài ra không cho thêm bất kì giả thiết nào khác. Điều này có nghĩa là phân phối mô hình phải thoả mãn các ràng buộc quan sát từ thực nghiệm và phải gần nhất với phân phối đều.

Entropy là độ đo về tính đồng đều hay tính không chắc chắn của một phân phối xác suất. Một phân phối xác suất có Entropy càng cao thì phân phối của nó càng đều. Độ đo Entropy điều kiện của một phân phối xác suất trên một chuỗi các trạng thái với điều kiện biết từ một chuỗi dữ liệu quan sát được tính như sau:

$$H(p) = -\sum_{x,y} \tilde{p}(x) p(y/x) \log p(y/x)$$

## 2.2 XÂY DỰNG MÔ HÌNH

Xem xét bài toán phân lớp, với Y là tập các lớp, X là tập các thông tin ngữ cảnh, là những thông tin quan trọng cần cho việc phân lớp văn bản vào lớp Y một cách chính xác.

Nhiệm vụ trong bài toán phân lớp là xây dựng một mô hình thống kê mà dự đoán chính xác lớp của văn bản bất kì. Mô hình như vậy chính là phương pháp ước lượng xác suất có điều kiện $p(y \mid x)$.

Mô hình Entropy cực đại cung cấp một phương pháp đơn giản để ước lượng xác suất có điều kiện $p(y \mid x)$ thông qua việc thống kê các thuộc tính quan trọng quan sát được từ tập dữ liệu huấn luyện.

### *2.2.1 Tập dữ liệu huấn luyện*

Để làm bài toán phân lớp trước tiên phải xây dựng tập dữ liệu huấn luyện T = {($x_1$, $y_1$), …, ($x_N$, $y_N$ )} trong đó {$x_1$, …, $x_N$} là tập các thông tin ngữ cảnh đã được gán nhãn tương ứng là tập các lớp {$y_1$, …, $y_N$ }.

Với một cặp ($x_i$, $y_i$), phân phối xác suất thực nghiệm của nó được tính bởi:

$$\tilde{p}(x_i, y_i) = \frac{1}{N} \times \text{số lần xuất hiện của } (x_i, y_i) \text{ trong tập dữ liệu mẫu}$$

Thông thường thì mỗi cặp($x_i$, $y_i$) không thể không xuất hiện trong tập mẫu mà nó sẽ xuất hiện ít nhất một lần.

### *2.2.2 Những thống kê, đặc trưng và ràng buộc*

Mục đích của chúng ta là xây dựng một mô hình thống kê của bài toán mà nó phát sinh xác suất $\tilde{p}(x, y)$ mẫu huấn luyện. Khối kiến trúc của mô hình này sẽ là một tập các thống kê của mẫu huấn luyện. Ví dụ khi xét bài toán phân loại bộ phim. Bộ phim được xếp vào một trong ba loại: good, not good, normal. Quan sát từ tập dữ liệu mẫu là 74 câu nhận xét đã được gán nhãn, ta có nhận xét như sau: nếu nhận xét có từ "failure" thì xác suất nhận xét đó thuộc loại "not good" là 80%. Đây chính là một thống kê.

Để biểu diễn sự kiện đó chúng ta có thể sử dụng hàm để biểu diễn như sau:

$$f(x, y) = \begin{cases} 1 & \text{If y = "failure"} \\ 0 & \end{cases}$$

Giá trị kỳ vọng của f liên quan tới phân phối thực nghiệm $\tilde{p}(x, y)$ chính là thống kê mà chúng ta đã nhắc tới. Chúng ta biểu diễn giá trị kỳ vọng này bởi:

$$\tilde{E}(f) = \sum \tilde{p}(x, y).f(x, y) \qquad \textit{với mọi cặp (x, y)} \quad (1)$$

Chúng ta có thể biểu diễn bất kỳ thống kê nào của mẫu huấn luyện như giá trị kỳ vọng của hàm nhị phân(f) thích hợp. Chúng ta gọi hàm đó là hàm đặc trưng hay đặc trưng. (Như vậy với các phân phối xác suất, chúng ta sẽ dùng ký hiệu và sử dụng hàm f(x, y) để biểu diễn giá trị của f với mỗi cặp (x, y) riêng biệt cũng như toàn bộ hàm f)

Khi chúng ta tìm hiểu về thống kê sẽ thấy sự hữu ích của nó, chúng ta có thể thấy được tầm quan trọng của nó bằng cách làm cho những gì có trong mô hình của chúng ta phù hợp với nó. Chúng ta làm điều này bằng các ràng buộc các giá trị kỳ vọng mà mô hình ấn định cho các hàm đặc trưng (f) tương ứng. Giá trị kỳ vọng của f quan hệ với xác suất mô hình p(y|x) như sau:

$$E(f) = \sum \tilde{p}(x).p(y \mid x).f(x,y) \qquad \textit{với mọi cặp (x, y)} \qquad (2)$$

Trong đó: p̃(x) là phân phối thực nghiệm của x trong mẫu huấn luyện. Chúng ta ràng buộc giá trị kỳ vọng này bằng với giá trị kỳ vọng của f trong mẫu huấn luyện:

$$E(f) = \tilde{E}(f) \qquad (3)$$

Từ (1), (2) và (3) ta có:

$$\sum_{x,y} \tilde{p}(x).p(y \mid x).f(x,y) = \sum_{x,y} \tilde{p}(x,y).f(x,y)$$

Chúng ta gọi (3) là phương trình ràng buộc hay đơn giản là ràng buộc. Bằng cách thu hẹp sự chú ý tới những xác suất mô hình p(y|x), như trong công thức (3), chúng ta loại trừ các mô hình được xem xét mà nó không thích hợp với mẫu huấn luyện dựa vào cách thông thường mà output của bài toán sẽ đưa ra đặc trưng f.

Tóm lại, chúng ta có được giá trị trung bình cho các thống kê tương ứng với các hiện tượng tồn tại trong dữ liệu mẫu, Ẽ(f), và cũng là giá trị

trung bình yêu cầu mà mô hình của bài toán đưa ra các hiện tượng đó (E(f) = Ẽ(f)).

Cần phân biệt rõ ràng 2 khái niệm về đặc trưng và ràng buộc: một đặc trưng là một hàm nhận giá trị nhị phân của cặp (x, y); một ràng buộc là một phương trình giữa giá trị kỳ vọng của hàm đặc trưng trong mô hình và giá trị kỳ vọng của nó trong dữ liệu huấn luyện.

### *2.2.3 Nguyên lý Entropy cực đại*

Giả thiết rằng chúng ta có *n* hàm đặc trưng $f_i$, nó quyết định những thống kê mà chúng ta cảm thấy là quan trọng trong quá trình mô hình hóa. Chúng ta muốn mô hình của chúng ta phù hợp với những thống kê đó. Vì vậy, chúng ta sẽ muốn p hợp lệ trong tập con **C** của **P** được định nghĩa bởi:

$$\boldsymbol{C} = \{p \in \boldsymbol{P} \mid E(f_i) = \tilde{E}(f_i) \quad for\ i \in \{1, 2, .., n\}\} \qquad (4)$$

Trong số các mô hình p € **C**, triết lý cực đại Entropy yêu cầu rằng chúng ta lựa chọn phân phối mà ngang bằng nhau nhất. Nhưng hiện tại chúng ta đối diện với câu hỏi rằng: ngang bằng nhau ở đây có nghĩa là gì?

Trong phạm vi toán học ngang bằng nhau của phân phối có điều kiện p(y|x) được cung cấp bởi Entropy có điều kiện:

$$H(p) = -\sum_{x,y} \tilde{p}(x).p(y \mid x).\log(p(y \mid x)) \qquad (5)$$

Entropy là bị chặn dưới bởi 0, Entropy của mô hình không có sự không chắc chắn nào, và chặn trên bởi log|**Y**|, Entropy của phân phối ngang bằng nhau trên toàn bộ các giá trị có thể |Y| của y. Với định nghĩa này, chúng ta đã sẵn sàng để biểu diễn nguyên lý của cực đại Entropy:

Để lựa chọn mô hình từ một tập **C** các phân phối xác suất được chấp nhận, lựa chọn mô hình $p^*$ € **C** với cực đại Entropy H(p):

$$p^* = \arg\max H(p) \qquad với\ p \in C \qquad (6)$$

Điều đó thể hiện rằng $p^*$ luôn luôn xác định; vì vậy, luôn luôn tồn tại một mô hình duy nhất $p^*$ với cực đại Entropy trong bất kỳ tập ràng buộc **C** nào.

### *2.2.4 Dạng tham số*

Từ nguyên lý Entropy cực đại ta có thể phát biểu lại rằng: tư tưởng chủ đạo của Entropy cực đại là ta phải xây dựng được một phân phối thoả mãn các ràng buộc và gần nhất với phân phối đều. Vấn đề đặt ra ở đây là làm thế nào để ta tối ưu được các ràng buộc, tức tìm ra được $p^* \in C$ làm cực đại $H(p)$. Trong những trường hợp đơn giản, chúng ta dễ dàng tìm ra mô hình phù hợp bằng các phương pháp giải tích. Tuy nhiên trong thực tế, số các ràng buộc là rất lớn và chồng chéo nhau. Vì vậy, chúng ta sẽ giải bài toán này theo một hướng tiếp cận khác.

Với mỗi một đặc trưng $f_i$, ta đưa vào một tham số $\lambda_i$ là một thừa số nhân Lagrange. Hàm Lagrange $\Lambda(p, \lambda)$ được định nghĩa như sau:

$$\Lambda(p,\lambda) = H(p) + \sum \lambda_i (E(f_i) - \tilde{E}(f_i)$$

Theo lý thuyết thừa số Lagrange, phân phối xác suất $p(y \mid x)$ làm cực đại độ đo Entropy $H(p)$ và thoả mãn tập ràng buộc C thì cũng làm cực đại hàm $\Lambda(p, \lambda)$ trên không gian phân phối xác xuất P. Gọi $P_\lambda$ là mô hình làm cực đại hàm Lagrange $\Lambda(p, \lambda)$, và $\Psi(\lambda)$ là giá trị cực đại.

$$P_\lambda = \arg\max \quad \Lambda(p,\lambda) \qquad \textit{với } p \in \boldsymbol{P}$$

$$\Psi(\lambda) = \Lambda(p,\lambda)$$

Ta gọi $\Psi(\lambda)$ là hàm đối ngẫu (**dual** function). Các hàm $p\lambda$, $\Psi(\lambda)$ đã được tính toán, chúng có công thức như sau:

$$p_\lambda(y \mid x) = (\frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i . f_i(x,y)))$$

$$\Psi(\lambda) = -\sum_x \tilde{p}(x).\log Z_\lambda(x) + \sum_i \lambda_i \tilde{E}(f_i)$$

Trong đó $Z_\lambda(x)$ là hằng số chuẩn hóa được quyết định bởi yêu cầu $\Sigma_y p_\lambda(y|x) = 1$ cho toàn bộ x:

$$Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y))$$

Cuối cùng thay cho việc phải tìm phân phối thoả mãn tập ràng buộc lớn và phức tạp làm cực đại độ đo Entropy, ta đưa về bài toán chỉ cần tìm tập tham số $\lambda^*$ làm cực đại hàm đối ngẫu $\Psi(\lambda)$ không còn ràng buộc.

Kết quả này có một ý nghĩa quan trọng. Khi đó, bất kì một thuật toán tìm cực đại $\lambda^*$ cho hàm $\Psi(\lambda)$ có thể sử dụng để tìm ra phân phối cực đại $* p$ của $H(p)$ thoả mãn $* p \in C$.

### *2.2.5 Mối quan hệ với cực đại Likelihood*

Log-likelihood $L_{\tilde{p}}(p)$ của phân phối thực nghiệm $\tilde{p}$ như được dự đoán trước bởi xác suất mô hình p được định nghĩa như sau:

$$L_{\tilde{p}}(p) = \log \prod_{x,y} p(y|x)^{\tilde{p}(x,y)} = \sum_{x,y} \tilde{p}(x, y).\log p(y|x)$$

Dễ dàng có thể kiểm tra được rằng hàm đối ngẫu *ψ(λ)* của phần trước chính là log-likelihood hàm số mũ của xác suất mô hình $p_\lambda$:

$$\Psi(\lambda) = L_{\tilde{p}}(p_\lambda)$$

Với cách giải thích này, kết quả của phần trước có thể được viết lại như sau: *mô hình $p^* \in \boldsymbol{C}$ với cực đại Entropy là mô hình trong đó họ tham số $p_\lambda(y|x)$ mà nó cực đại likelihood của xác suất mẫu huấn luyện $\tilde{p}$.*

Kết quả này giúp làm tăng thêm tính đúng đắn cho nguyên lý cực đại Entropy: khi quan niệm việc lựa chọn xác suất mô hình p* trên cơ sở cực đại Entropy là không đủ sức thuyết phục, điều xảy ra với cùng một xác suất

p* là một mô hình mà nó, trong số toàn bộ các mô hình của cùng một dạng tham số, có thể là sự miêu tả tốt nhất cho mẫu huấn luyện.

### *2.2.6 Các thuật toán ước lượng tham số*

Có nhiều thuật toán dùng để ước lượng tham số, điển hình là các thuật toán GIS, IIS, L-BFGS. Trong khoá luận này, chúng tôi xin giới thiệu thuật toán L-BFGS là thuật toán ước lượng tập tham số hiệu quả nhất hiện nay.

Cho tập dữ liệu huấn luyện: $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

Phân phối mũ:

$$p_\lambda(y \mid x) = \left(\frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i . f_i(x, y)\right)\right)$$

Huấn luyện mô hình Entropy cực đại chính là ước lượng tập trọng số $\lambda = \{ \lambda_i, .., \lambda_k\}$để phân phối mũ ở trên đạt cực đại cao nhất.

Thuật toán L-BFGS là phương pháp giới hạn bộ nhớ cho phương pháp quasi-Newton(*Limited memory BFGS*). Phương pháp này cung cấp khả năng tối ưu hàng triệu tham số với tốc độ rất nhanh, vì vậy trong các nghiên cứu mới đây nó được đánh giá là hiệu quả hơn các phương pháp khác.

Viết lại hàm log-likelihood khi thay $p_\lambda( y \mid x)$ từ () vào ():

$$L(p_\lambda) =$$

$$\sum_{i=1}^{N} \sum_{j=1}^{k} \lambda_j f_j\ (x_i, y_i) - \sum_{i=1}^{N} log \sum_{a \epsilon A} \exp\ (\sum_{j=1}^{k} \lambda_j f_j(y, x_i))$$

Tư tưởng của thuật toán là sử dụng phương pháp leo đồi tìm kiếm cực đại toàn cục. Vì bề mặt của hàm $L(p_\lambda)$ là lồi nên hoàn toàn có thể thực hiện được điều này. Các thủ tục lặp được sử dụng để tiến gần đến tối ưu toàn cục của hàm $L(p_\lambda)$. Tại mỗi bước lặp ta tìm vec-tơ gradient nào có hướng tiến tới cực đại toàn cục nhất. Trên bề mặt của một hàm lồi, vec-tơ

gradient thoả mãn điều kiện đó sẽ có giá trị bằng $\vec{0}$. Với mỗi một vec-tơ gradient($\frac{\partial L(p_\lambda)}{\partial \lambda_1}, \ldots, \frac{\partial L(p_\lambda)}{\partial \lambda_N}$) hiện tại xác định cho ta một tập các trọng số.

Thành phần thứ i của vecto gradient của $L(p_\lambda)$ là:

$$\frac{\partial L(p_\lambda)}{\partial \lambda_1} = \sum_{j=1}^{N} f_j(x_j, y_j) - \sum_{j=1}^{N} \frac{\sum_{y \epsilon Y} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(y, x_j)\right) f_{j(y,x_j)}}{\sum_{y \epsilon Y} \exp\left(\sum_{i=1}^{n} \lambda_i f_i(y, x_j)\right)}$$

$$= \sum_{j=1}^{N} f_i(x_j, y_j) - \sum_{j=1}^{N} \sum_{y \epsilon Y} p_\lambda\left(y \middle| x_j\right) f_i(y, x_j)$$

$$= E_p f_i(x, y) - E_{p\lambda} f_i(x, y)$$

Trong mỗi bước lặp thủ tục L-BFGS yêu cầu giá trị hiện tại của $L(p_\lambda)$ và vecto gradient hiện tại. Sau đó nó tính toán để cập nhật giá trị mới cho tập tham số $\{ \lambda_i, \ldots, \lambda_k \}$. Cuối cùng ta thu được tập trọng số tối ưu $\{\lambda_i^*, \ldots, \lambda_k^*\}$ sau một số hữu hạn các bước lặp.

$$P(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

## 2.3 BÀI TOÁN PHÂN LỚP QUAN ĐIỂM SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY MAXIMUM ENTROPY CỰC ĐẠI

Để thực hiện phương pháp học máy này với dữ liệu tài liệu, chúng ta sử dụng các đặc trưng N-gram.

Đặt $\{ f_1, \ldots, f_m \}$ là một bộ được xác định trước của đặc trưng m có thể xuất hiện trong một tài liệu, ví dụ bao gồm từ "still" hoặc "really stinks".

Đặt $n_i(d)$ là số lần xuất hiện của $f_i$ trong tài liệu d.

Tiếp theo, mỗi vector đặc trưng đại diện cho 1 văn bản d:

$$\overrightarrow{d} := (n_1(d), n_2(d), \ldots, n_m(d)).$$

Phân loại Maximum Entropy (MaxEnt, or ngắn gọn là: ME) là một kỹ thuật thay thế hiệu quả đã được chứng minh trong một ứng dụng xử lý ngôn ngữ tự nhiên cho thấy rằng đôi khi nhưng không phải lúc nào cũng thực hiện tốt hơn Naive Bayes về phân loại văn bản chuẩn. Nó ước tính P(c/d) theo dạng số mũ như sau :

$$P_{ME}(c/d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right),$$

Trong đó: Z(d) là một hàm chuẩn hóa

$F_{i,c}$ là một đặc trưng / lớp hàm

$f_i$: đặc trưng

c: lớp

$$F_{i,c}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ và } c' = c \\ 0 & \text{khác} \end{cases}.$$

Cho ví dụ, một đặc trưng đặc biệt / lớp hàm có thể chạy khi và chỉ khi <bigram> "*still hate*" xuất hiện và quan điểm của tài liệu được giả thuyết là tiêu cực. Điều quan trọng là không giống như Naive Bayes, MaxEnt không giả định về mối quan hệ giữa các đặc trưng, và vì vậy có thể có khả năng thực hiện tốt hơn khi giả định điều kiện độc lập không được đáp ứng.

$\lambda_{i,c}$ là các tham số của đặc trưng; kiểm tra định nghĩa của $P_{ME}$ cho thấy rằng $\lambda_{i,c}$ lớn có nghĩa $f_i$ được xét đến là 1 chỉ số quan trọng của lớp c. Các giá trị của tham số được thiết lập để cực đại hóa Entropy trong sự phân loại đưa ra tùy theo điều kiện ràng buộc đó mà các giá trị được kỳ vọng của lớp các hàm, đối với bản mẫu bằng những giá trị kỳ vọng của chúng, đối với dữ liệu huấn luyện: triết lý cơ bản đó là chúng ta nên chọn mô hình tạo ra các giả định ít nhất về dữ liệu trong khi "*still*" vẫn còn phù hợp với nó, làm ý nghĩa trực quan hơn.

# Chương 3: THỰC NGHIỆM

## 3.1 DỮ LIỆU THỬ NGHIỆM

Trong đồ án này 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 nguồn của chúng tôi là bản lưu trữ Internet movie Database của rec.arts.movies.reviews newgroup. Tập dữ liệu này sẽ có sẵn tại:

*http://www.cs.cornell.edu/people/pabo/movie-review-data /*. Các dữ liệu này đã được 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>.

***Phát biểu bài toán:*** bài toán phân lớp quan điểm được phát biểu như sau:

Cho một file văn bản chứa các câu được biểu diễn: { $f_1$,. . . , $f_m$ } là một bộ được xác định trước của m đặc trưng xuất hiện trong một tài liệu.

$n_i(d)$ là số lần xuất hiện của $f_i$ trong tài liệu d.

Tiếp theo, mỗi vector đặc trưng đại diện cho 1 văn bản d:

$$\overrightarrow{d} := (n_1(d), n_2(d),. . . , n_m(d)).$$

văn bản d chứa quan điểm tiêu cực hay tích cực. Ta thực hiện phân lớp cho văn bản bằng cách gán nhãn: 1 hoặc -1 cho từng câu trong văn bản.

1: tương ứng với văn bản chứa quan điểm tích cực.

-1: tương ứng với văn bản chứa quan điểm

**Input:** Cho một tập hợp các câu văn bản đánh giá *D* có ý kiến (hoặc tình cảm) về một đối tượng.

**Output:** Mỗi câu được gán nhãn phân loại là câu chứa quan điểm tích cực hay câu chứa quan điểm tiêu cực.

## 3.2 CÔNG CỤ SỬ DỤNG

### *3.2.1 Công cụ sinh SRIML*

✓ ***Giới thiệu***: SRILM– The SRI Language Modeling Toolkit là một bộ công cụ để xây dựng và áp dụng mô hình ngôn ngữ thống kê (LMS), chủ yếu là để sử dụng trong nhận dạng giọng nói, gắn thẻ thống kê và phân đoạn và một máy dịch thuật.

✓ ***Thành phần:***

➢ Một tập hợp các thư viện lớp C + + thực hiện các mô hình ngôn ngữ, hỗ trợ stuctures dữ liệu và các chức năng tiện ích linh tinh.

➢ Một tập hợp các chương trình thực thi được xây dựng trên đầu trang của các thư viện để thực hiện nhiệm vụ tiêu chuẩn như các LMS đào tạo và thử nghiệm chúng gắn thẻ, dữ liệu hoặc phân chia văn bản, vv

✓ ***Cách thực hiện huấn luyện mô hình ngôn ngữ***

**ngram –count –order*n* -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 Công cụ phân lớp dữ liệu Maxent*

Công cụ mã nguồn mở Maxent của tác giả Le Zhang tại Centre for Speech Technology Research, University of Edinburgh.

http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Bộ công cụ Maximum Entropy cung cấp một tập hợp các công cụ và thư viện để xây dựng mô hình Entropy tối đa (Maxent) trên nền Python hay C++.

Gồm 3 thành phần chính:

- Huấn luyện mô hình
- Kiểm thử dữ liệu
- Tính toán độ chính xác của mô hình phân lớp + gán nhãn phân loại cho câu chứa quan điểm.

***Cách thực hiện:***

*Bước 1*: Huấn luyện dữ liệu

***maxent -m <model><trainFile>***

Trong đó:

- <***trainFile***> là tên file dữ liệuhuấn luyên đã được gán nhãn
- -m: thiết lập tên mô hình

- <***model***> tên mô hình người dùng được huấn luyện

*Bước 2:* Quá trình test

***maxent -p -m <model><testFile>-o <outputFile>***

Trong đó:

- -p: dự đoán độ chính xác của quá trình huấn luyện
- <***model***>: tên mô hình đã được huấn luyện ở bước trên
- **<testFile:**tên file dữ liệu test
- -o: dự đoán tên file kết quả
- <***outputFile***>: tên file kết quả

## 3.3 *KẾT QUẢ THỬ NGHIỆM*

Thực hiện trên bộ dữ liệu polariry review 1.0

***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 (mặc định là 10) sẽ bị loại bỏ. Bước cuối cùng được thực hiện sau khi đã gán các trọng số cho từng đặc trưng. Tại bước này, những đặc trưng nào không làm tăng Entropy của mô hình thì sẽ bị loại bỏ.

*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 Maxent.

Để thực hiện phân lớp tài liệu quan điểm, chúng tôi chia tập dữ liệu thành hai tập con là tập huấn luyện và tập test.

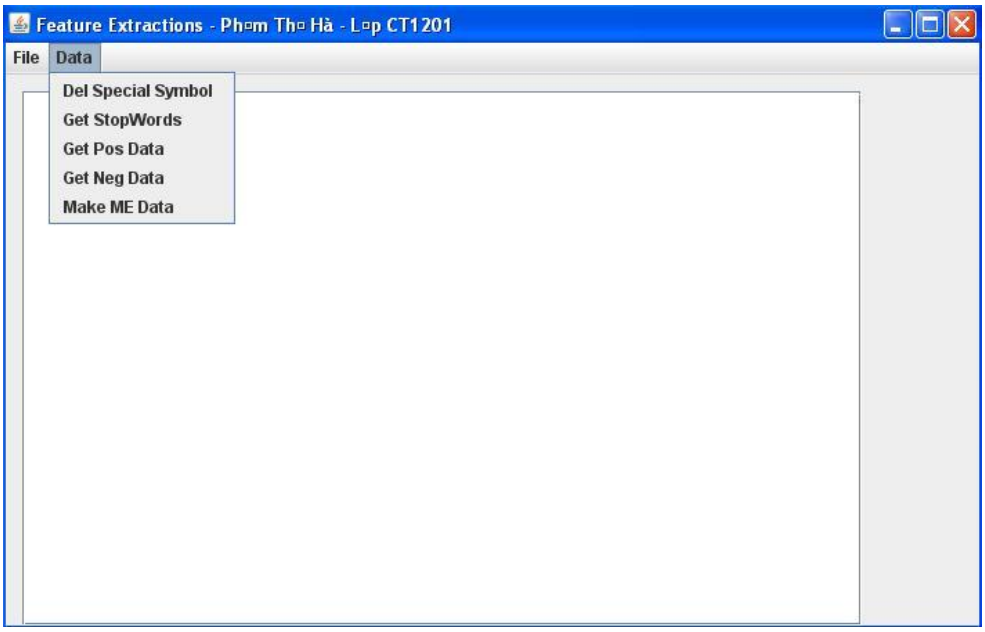
Tập huấn luyện gồm có 550 nhận xét tích cực và 550 nhận xét tiêu cực.

Tập test gồm có 150 nhận xét tích cực và 150 nhận xét tiêu cực.

Kết quả thực hiện phân lớp Maximum Entropy với các đặc trưng Uni-gram và Bigram như sau:

| Đặc trưng | Uni-gram | Bi-gram |
|---|---|---|
| Độ chính xác | 83,2776 % | 79,5987 % |

Giao diện chính của chương trình:




**Một số hình ảnh khi chạy chương trình:**


*Hiển thị dữ liệu 1 trong 3 file sau khi chạy chức năng* ***Get Neg Data***


*Hiển thị dữ liệu 1 trong 3 file sau khi chạy chức năng* ***Get Pos Data***

```
C:\Documents and Settings\nghia>cd\

C:\>D:

D:\>maxent.exe
maxent 20041229

Purpose:
  A command line utility to train (test) a maxent model from a file.

Usage: maxent [OPTIONS]... [FILES]...
   -h          --help              Print help and exit
   -V          --version           Print version and exit
   -v          --verbose           verbose mode (default=off)
   -mSTRING    --model=STRING      set model filename
   -b          --binary            save model in binary format (default=off)
   -oSTRING    --output=STRING     prediction output filename
               --detail            output full distribution in prediction mode (def
ault=off)
   -iINT       --iter=INT          iterations for training algorithm (default='30')

   -gFLOAT     --gaussian=FLOAT    set Gaussian prior, disable if 0 (default='0.0')

   -cINT       --cutoff=INT        set event cutoff (default='1')
               --heldout=STRING    specify heldout data for training
   -r          --random            randomizing data in cross validation (default=of
f)
               --nommap            do not use mmap() to read data (slow) (default=o
ff)

   Group: MODE
   -p          --predict           prediction mode, default is training mode
   -nINT       --cv=INT            N-fold cross-validation mode (default='0')

   Group: Parameter Estimate Method
               --lbfgs             use L-BFGS parameter estimation (default)
               --gis               use GIS parameter estimation

D:\>maxent /Ha/outData_Ngram/uTrain.txt -m MO1

D:\>maxent -p -m MO1 /Ha/outData_Ngram/uTest.txt -o out_u.txt
Accuracy: 83.2776% (249/299)

D:\>maxent /Ha/outData_Ngram/bTrain.txt -m MO2

D:\>maxent -p -m MO2 /Ha/outData_Ngram/bTest.txt -o out_b.txt
Accuracy: 79.5987% (238/299)

D:\>
```


*Hình ảnh file kết quả sau khi chạy Maxent*

# KẾT LUẬN

Trong quá trình làm khóa luận, em đã tìm hiểu được mô hình Entropy cực đại, một số khía cạnh về phân lớp quan điểm và các vấn đề đặt ra với bài toán này.

Đồ án tìm hiểu về tiếp cận phân lớp quan điểm sử dụng trích chọn đặc trưng n-gram và áp dụng phương pháp học máy Maximum Entropy.

Chương trình thực nghiệm cũng đã thử nghiệm trên bộ dữ liệu với 700 câu tích cực và 700 câu tiêu cực và sử dụng phương pháp học máy có giám sát Maximum Entropy để phân lớp với cách chọn đặc trưng là sử dụng uni-gram và bi-gram.

Hướng nghiên cứu tiếp theo, em sẽ tiếp tục thử nghiệm với các phương pháp phân loại khác và nghiên cứu các phương pháp chọn đặc trưng để phân loại câu quan điểm hiệu quả.

Trong một khoảng thời gian có hạn, nên khi trình bày các vấn đề em đã nghiên cứu được không tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu của thầy cô và các bạn.

Em xin chân thành cảm ơn!

# TÀI LIỆU THAM KHẢO


1. Bo Pang and Lillian Lee và Shivakumar Vaithyanathan. **Thumbs up Sentiment Classification using Machine Learning Techniques.**

2. **Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews.** Nipun Mehra and Shashikant Khandelwal and Priyank Patel.

3. 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.

4. Nguyễn Thùy Linh về "**Phân lớp tài liệu web độc lập ngôn ngữ**". Khóa luận tốt nghiệp đại học hệ chính quy, ngành công nghệ thông tin, trường Đại học Quốc gia Hà Nội.

5. http://www.cs.cornell.edu

6. http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

7. ftp://ftp.cs.cornell.edu/pub/smart/english.stop

8. http://www.speech.sri.com/projects/srilm/download.html