BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM**

---------------------------



**VÕ MINH QUÂN**

**ỨNG DỤNG KHAI THÁC DỮ LIỆU VÀO LĨNH VỰC GIÁO DỤC**

**LUẬN VĂN THẠC SĨ**

Chuyên ngành : Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 06 năm 2020

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM**

---------------------------



**VÕ MINH QUÂN**

**ỨNG DỤNG KHAI THÁC DỮ LIỆU VÀO LĨNH VỰC GIÁO DỤC**

**LUẬN VĂN THẠC SĨ**

Chuyên ngành : Công nghệ thông tin

Mã số ngành: 60480201

**CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS Lê Thị Ngọc Thơ**

TP. HỒ CHÍ MINH, tháng 06 năm 2020

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM**  Cán bộ hướng dẫn khoa học : TS Lê Thị Ngọc Thơ  Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM  ngày … tháng 06 năm 2020  Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:  *(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)*       |  |  |  | | --- | --- | --- | | **TT** | **Họ và tên** | **Chức danh Hội đồng** | | 1 |  | Chủ tịch | | 2 |  | Phản biện 1 | | 3 |  | Phản biện 2 | | 4 |  | Ủy viên | | 5 |  | Ủy viên, Thư ký |     Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được  sửa chữa (nếu có).  Chủ tịch Hội đồng đánh giá LV |

|  |  |
| --- | --- |
| TRƯỜNG ĐH CÔNG NGHỆ TP. HCM  **VIỆN ĐÀO TẠO SAU ĐẠI HỌC** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

*TP. HCM, ngày*..*… tháng…*.. *năm 20*..*…*

**NHIỆM VỤ LUẬN VĂN THẠC SĨ**

Họ tên học viên: VÕ MINH QUÂN Giới tính: Nam

Ngày, tháng, năm sinh: 16/11/1995 Nơi sinh: Vĩnh Long

Chuyên ngành: Công nghệ thông tin MSHV: 1741860036

**I- Tên đề tài:**

ỨNG DỤNG KHAI THÁC DỮ LIỆU VÀO LĨNH VỰC GIÁO DỤC

**II- Nhiệm vụ và nội dung:**

**III- Ngày giao nhiệm vụ:** *20/03/2019*

**IV- Ngày hoàn thành nhiệm vụ:** dd/mm/yyyy

**V- Cán bộ hướng dẫn:** **TS Lê Thị Ngọc Thơ**

**CÁN BỘ HƯỚNG DẪN KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký) (Họ tên và chữ ký)

**LỜI CAM ĐOAN**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này   
đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

**Học viên thực hiện Luận văn**

**Võ Minh Quân**

**LỜI CẢM ƠN**

Trải qua một thời gian dài tìm hiểu và nỗ lực nghiên cứu cuối cùng tôi đã hoàn thành được luận văn thạc sĩ với đề tài: ***“*Ứng dụng khai thác dữ liệu vào lĩnh vực giáo dục*”.***

Để hoàn thành luận văn thạc sĩ này, lời đầu tiên tôi xin chân thành cảm ơn quý thầy/cô khoa Công nghệ thông tin trường Đại Học Công Nghệ TP HCM những người đã trực tiếp giảng dạy, truyền đạt những kiến thức bổ ích cho tôi trong suốt thời gian học tập tại trường, đó chính là những nền tảng kiến thức cơ bản, là những hành trang vô cùng quý giá góp phần xây dựng nên luận văn này.

Và đặc biệt tôi xin gửi một lời cảm ơn sâu sắc đến ***Ts Lê Thị Ngọc Thơ***, cô đã là người trực tiếp hướng dẫn tôi trong suốt quá trình học tập và nghiên cứu xây dựng luận này. Cô đã tận tình quan tâm, giúp đỡ tôi trong quá trình học tập, giải đáp những thắc mắc kịp thời và rõ ràng trong suốt quá trình làm luận văn. Nhờ đó tôi mới có thể hoàn thành được luận văn này theo kịp tiến độ.

Tôi xin cảm ơn tập thể lớp 17SCT21, trường Đại học Công nghệ TP. Hồ Chí Minh đã cung cấp, hỗ trợ nguồn tài liệu, đóng góp ý kiến trong quá trình học tập nghiên cứu luận văn này.

Và cuối cùng cũng xin bày tỏ lòng biết ơn sâu sắc đến cha mẹ, những người đã sinh thành, dưỡng dục tôi nên người và tạo điều kiện để đạt được kết quả ngày hôm nay .

Tuy có nhiều cố gắng trong quá trình thực hiện khóa luận không thể tránh khỏi những thiếu sót, tôi mong được sự góp ý từ quý thầy cô cũng như tất cả bạn bè để đạt kết quả hoàn thiện hơn.

Một lần nữa tôi xin chân thành cảm ơn.

**Võ Minh Quân**

**TÓM TẮT**

+ Họ và tên học viên: Võ Minh Quân.

+ Chuyên ngành: Công nghệ thông tin Lớp: 17SCT21

* Cán bộ hướng dẫn: TS Lê Thị Ngọc Thơ.
* Tên đề tài: Ứng dụng khai thác dữ liệu vào lĩnh vực giáo dục.

Ngày nay cùng với sự phát triển mạnh mẽ của công nghệ mọi nguồn dữ liệu trong các lĩnh vực dần được số hóa và mang lại cho chúng ta một nguồn tài nguyên phong phú để có thể tận dụng khai thác. Vì vậy những năm gần đây việc phân tích khai thác dữ liệu ngày được chú trọng và phát triển hơn, các đề tài nghiên cứu ra đời ngày càng nhiều. Tuy nhiên số lượng các nghiên cứu tại Việt Nam vẫn còn hạn chế đặc biệt là ở lĩnh vực giáo dục, vì vậy tôi quyết định chọn hướng nghiên cứu này cho luận văn của mình.

Luận văn này nghiên cứu về kỹ thuật khai thác dữ liệu bằng các phương pháp khác nhau và ứng dụng vào bài toán thực tế trong lĩnh vực giáo dục. Cụ thể tôi đã tập trung nghiên cứu về các khái niệm xử lý văn bản, tóm tắt văn bản, phân loại văn bản đồng thời tìm hiểu những phương pháp phân tích văn bản, phân lớp văn bản khác nhau để áp dụng vào cụ thể bài toán của luận văn. Đặc biệt là tập trung nghiên cứu về kỹ thuật phân tích xử lý văn bản và phân lớp dữ liệu.

Phương pháp tiếp cận bài toán phân tích xử lý văn bản và phân lớp dữ liệu tôi trải qua các bước sau:

Bước 1: Thực hiện việc tách câu từ từ các tập dữ liệu.

Bước 2: Mô hình hóa các câu từ thành vector.

Bước 3: Chạy huấn luyện và phân lớp dữ liệu qua các phương pháp phân lớp khác nhau.

Với cách tiếp cận trên của tôi và áp dụng trực tiếp vào bài toán phân tích ý kiến khảo sát đánh giá giảng viên tại trường Đại học Công nghệ TP.HCM tôi đã thu về một số kết quả nhất định. Sau quá trình thực nghiệm bài toán tôi đã xây dựng một bộ phân lớp dữ liệu đánh giá giảng viên đáng tin cậy với độ chính xác lên đến ~75% và rút ra được những nhận xét về ưu nhược điểm của các phương pháp phân lớp.

Với kết quả này luận văn đã đóng góp một bộ phân lớp dữ liệu mới trong lĩnh vực giáo dục cụ thể là bộ phân lớp ý kiến đánh giá giảng viên và có thể áp dụng vào các ứng dụng thực tế. Bên cạnh đó luận văn còn là nguồn tài liệu tham khảo về ưu nhược điểm của một số phương pháp phân lớp trên cùng tập dữ liệu qua đó quyết định lựa chọn phương pháp phù hợp.

Luận văn này bao gồm 5 chương – trình bày chi tiết các ý tưởng, phương thức thực hiện, các thực nghiệm và kết luận cũng như hướng phát triển tiếp theo cho đề tài.

**ABSTRACT**

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 14](#_Toc40950030)

[1.1 Giới thiệu 14](#_Toc40950031)

[1.2 Tính cấp thiết luận văn 15](#_Toc40950032)

[1.3 Mục tiêu luận văn 15](#_Toc40950033)

[1.4 Nội dung nghiên cứu 15](#_Toc40950034)

[1.5 Phương pháp nghiên cứu 16](#_Toc40950035)

[1.6 Bố cục luận văn 16](#_Toc40950036)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 18](#_Toc40950037)

[2.1 Phân tích ý kiến 18](#_Toc40950038)

[2.2 Phân tích cảm xúc 19](#_Toc40950039)

[*2.2.1* *Phương pháp phân lớp Naïve Bayes* 20](#_Toc40950040)

[*2.2.2* *Phương pháp phân lớp SVM (support vector machines).* 20](#_Toc40950041)

[*2.2.3* *Phương pháp K-Nearest Neighbor (k-NN)* 20](#_Toc40950042)

[*2.2.4* *Phương pháp Phương pháp Linear Least Square Fit (LLSF)* 20](#_Toc40950043)

[*2.2.5* 20](#_Toc40950044)

[2.3 Phân lớp câu chủ quan 21](#_Toc40950045)

[2.4 Các mô hình biểu diễn văn bản 21](#_Toc40950046)

[*2.4.1* *Mô hình logic* 21](#_Toc40950047)

[*2.4.2* *Mô hình phân tích cú pháp* 22](#_Toc40950048)

[*2.4.3* *Mô hình không gian vector* 23](#_Toc40950049)

[*2.4.4* *Mô hình Boolean* 25](#_Toc40950050)

[*2.4.5* *Mô hình tần số từ khóa (TF – Term Frequency)* 26](#_Toc40950051)

[*2.4.6* *Mô hình nghịch đảo tần số văn bản (IDF – Inverse Document Frequency)* 26](#_Toc40950052)

[*2.4.7* *Mô hình TF - IDF* 26](#_Toc40950053)

[2.5 26](#_Toc40950054)

[2.6 Phân lớp câu chủ quan 26](#_Toc40950055)

[TÀI LIỆU THAM KHẢO 30](#_Toc40950056)

**DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Viết tắt** | **Tiếng Anh** | **Tiếng Việt** |
| 1 | CSDL | Database | Cơ sở dữ liệu |
| 2 | SVM | Support Vector Machines | Máy vector hỗ trợ |
| 3 | NB | Naïve Bayes |  |
| 4 | KNN | k-Nearest Neighbors | k-láng giềng gần |

**DANH MỤC CÁC BẢNG**

**DANH MỤC CÁC HÌNH**

# CHƯƠNG 1: GIỚI THIỆU

# Giới thiệu

Như chúng ta đã biết từ ngày xa xưa cho đến hiện nay trong nhiều lĩnh vực của xã hội việc thu thập ý kiến, cảm nhận, phản hồi đánh giá của con người là một việc rất phổ biến mà dựa vào đó chúng ta có thể để đưa ra những đánh giá, nhận xét liên quan. Ở những giai đoạn trước khi công nghệ chưa phát triển hình thức này được diễn ra dưới dạng hòm thư góp ý, lấy ý kiến trực tiếp.v.v. Trong những năm gần đây với sự bùng nổ của ngành công nghệ thông tin công việc này ngày càng được chú trọng và số hóa nhiều hơn.

Một vài nguồn tài nguyên phổ biến cho việc thu thập và sử dụng ý kiến phản hồi được kể ra như dưới đây:

* Kinh nghiệm cá nhân và ý kiến về bất cứ điều gì trong đánh giá, diễn đàn v.v.
* Nhận xét về bài viết, vấn đề, chủ đề, bài đánh giá, v.v.
* Thông tin đăng tại các trang web mạng xã hội, ví dụ: Facebook.
* Đánh giá về các dịch vụ, sản phẩm.

Vậy tại sao những ý kiến này lại quan trọng đến như vậy? Những luận điểm dưới đây sẽ giải đáp những vấn đề này:

* “Ý kiến” là những yếu tố ảnh hưởng quan trọng đến hành vi của một người.
* Những ý kiến đánh giá là một phần quan trọng để đo lường và đánh giá chất lượng sản phẩm hay dịch vụ.
* Thực tế khi chúng ta cần đưa ra quyết định, chúng ta thường tìm kiếm ý kiến của người khác. Với cá nhân sẽ tìm kiếm ý kiến từ bạn bè và gia đình, còn với tổ chức sử dụng khảo sát ý kiến, tư vấn.

Những ứng dụng từ việc phân tích ý kiến cũng được áp dụng rộng rãi trong nhiều lĩnh vực. Đối với các doanh nghiệp và tổ chức, việc phân tích ý kiến hỗ trợ việc cung ứng ra thị trường các sản phẩm phù hợp nhu cầu và xu hướng. Đối với cá nhân, việc phân tích ý kiến có thể hỗ trợ người dùng trong quá trình ra quyết định sử dụng dịch vụ, thu thập các ý kiến liên quan đến môi trường xã hội xung quanh.

Tuy nhiên hiện nay việc phân tích xử lý các dữ liệu này phần lớn còn được thực hiện một cách thủ công dưới sự đánh giá trực tiếp từ con người. Vì vậy những hệ thống phân tích ý kiến tự động và đưa ra những tổng hợp đánh giá là một nhu cầu cần thiết sẽ mang lại nhiều giá trong nhiều lĩnh vực. Trong lĩnh vực giáo dục việc áp dụng một hệ thống phân tích ý kiến dự đoán tự động những ý kiến đánh giá của học sinh, sinh viên về chất lượng giảng viên trong các khóa học, chương trình đào tạo sẽ giúp tiết kiệm một lượng lớn nguồn nhân lực cũng như thời gian đánh giá.

# Tính cấp thiết luận văn

Sau một thời gian tìm hiểu và phân tích chúng tôi nhận thấy việc thu thập ý kiến đánh giá chất lượng giảng dạy của sinh viên trong mỗi học kì ở Trường Đại học Công nghệ TP.HCM hiện nay là một bài toán thực tế và có thể áp dụng được mô hình phân tích và đánh giá ý kiến. Với một lượng dữ liệu rất lớn về việc ý kiến đánh giá của sinh viên trong mỗi học kì thì việc tổng hợp và đánh giá thủ công thông qua con người sẽ tốn rất nhiều thời gian và chi phí.

Vì vậy luận văn này sẽ thực hiện nghiên cứu và áp dụng phân tích, tổng hợp các ý kiến đánh giá một cách tự động. Mục tiêu của nghiên cứu này là giúp rút ngắn thời gian thực hiện đánh giá, phân tích bên cạnh đó sẽ hỗ trợ đánh giá chất lượng được khách quan hơn.

# Mục tiêu luận văn

Mục tiêu nghiên cứu chính của luận văn là tìm hiểu về các phương pháp phân tích ý kiến và phân lớp dữ liệu. Bên cạnh đó nghiên cứu cũng sẽ so sánh độ hiệu quả giữa các phương pháp phân lớp dữ liệu thông qua bài toán phân tích ý kiến đánh giá giảng viên.

Đối với bài toán phân tích ý kiến đánh giá của sinh viên về chất lượng giảng dạy tại Trường Đại học Công nghệ TP.HCM tôi dự kiến tạo được một hệ thống phân tích các ý kiến thu thập được một cách tự động, xác định được cụ thể ý kiến là đánh giá tích cực hay tiêu cực.

Đưa ra các kết luận, đánh giá về kết quả đạt được đồng thời cũng nêu ra phương hướng để giải quyết các vấn đề còn tồn tại.

Ngoài ra luận văn này có thể phát triển thêm ở việc xác định khía cạnh đánh giá của ý kiến, hướng phát triển này phụ thuộc vào độ hiệu quả của việc đánh giá ý trước trước đó.

# Nội dung nghiên cứu

Dựa vào các mục tiêu đã xác định luận văn sẽ tiến hành nghiên cứu các nội dung sau:

* Nghiên cứu về phân lớp chủ quan về phân lớp cảm nghĩ.
* Nghiên cứu về tóm tắt ý kiến.
* Nghiên cứu về phương pháp phân lớp dữ liệu.
* Nghiên cứu về phân loại ý kiến dựa trên học không giám sát.
* Xây dựng bộ phân lớp dữ liệu đánh giá giảng viên.
* So sánh độ hiệu quả của bộ phân lớp qua các phương pháp khác nhau.
* Kết luận đưa ra các đánh giá.

Thực nghiệm và đánh giá trên CSDL khảo sát sinh viên năm học 2016-2017.

# Phương pháp nghiên cứu

Tìm hiểu các tài liệu về phân tích ý kiến, cảm xúc thông qua các từ khóa phổ biến như: opinion mining, data mining opinion, data mining and education, v.v.

Tìm hiểu các phương pháp liên quan đến khai thác văn bản, ý kiến, phân lớp dữ liệu, học giám sát, học không giám sát so sánh độ hiệu quả giữa các phương pháp thông qua các ứng dụng thực tế đã.

Tìm hiểu các kỹ thuật xử lý văn bản, phân lớp văn bản lựa chọn ra các phương pháp phù hợp để áp dụng vào bài toán của luận văn.

Cài đặt các thuật toán của các phương pháp đã nghiên cứu

Chạy thực nghiệm các dữ liệu đánh giá giảng viên trên các thuật toán đã cài đặt, ghi nhận kết quả và đánh giá nhận xét.

# Bố cục luận văn

Luận văn sẽ dự kiến thực hiện 5 chương:

* Chương 1: Giới thiệu tổng quan về đề tài, tính cấp thiết luận văn, mục tiêu nghiên cứu, nội dung nghiên cứu, phương pháp nghiên cứu.
* Chương 2: Trình bày cơ sở lý thuyết về phân tích ý kiến, phân loại cảm xúc, phân lớp câu chủ quan, các mô hình biểu diễn văn bản, tóm tắt văn bản, từ vựng văn bản. Tìm hiểu các nghiên cứu đã có về phân tích ý kiến, phân loại cảm xúc, phân lớp câu chủ quan.
* Chương 3: Phương pháp thực hiện gồm thu thập dữ liệu và tiền xử lý dữ liệu bằng các phương pháp hiệu quả, gán nhãn dữ liệu lựa chọn theo quy tắc và lựa chọn các phương pháp phân tích ý kiến, phân lớp cảm xúc để áp dụng.
* Chương 4: Thực nghiệm và đánh giá gồm thu thập dữ liệu từ nguồn dữ liệu khảo sát sinh viên học kỳ 2 năm học 2016-2017 tại trường Đại học Công Nghệ Tp.HCM tiến hành trích xuất và tiền xử lý, chuẩn bị môi trường thiết lập các thuật toán thực nghiệm, trình bày về các công cụ cần cho thực nghiệm, cài đặt các thuật toán đã tìm hiểu trên môi trường đã chuẩn bị, chạy thực nghiệm dữ trên các phương pháp khác nhau, trình bày kết quả thực nghiệm trên tập dữ liệu đánh giá sinh viên trên các phương pháp khác nhau và cuối cùng là phân tích so sánh kết quả thu được thông qua các phương pháp.
* Chương 5: Kết luận và hướng phát triển.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

# Phân tích ý kiến

Phân tích ý kiến hiện nay đang trở thành một trong những lĩnh vực được chú trọng và phát triển. Rất nhiều nghiên cứu trong lĩnh vực này đã ra đời mang lại một cái nhìn phong phú đa chiều cho lĩnh vực.

Như vào năm 2006 Jindal và Liu đã đưa ra một nhận xét thì ý kiến thường xuyên có 2 loại là: **cảm xúc** và **ý kiến** [1].

Trong khi đó Hu và Liu thì lại cho rằng một ý kiến có cấu trúc bao gồm **thực thể** và **khía cạnh** [2].

Sau đó vào năm 2010 để phân tích ý kiến Liu đã đưa các ý kiến về theo một cấu trúc gồm năm thành phần [3]:

Trong đó:

* là một thực thể đích.
* là một khía cạnh/tính năng của thực thể .
* là giá trị cảm xúc của ý kiến từ người giữ ý kiến về tính năng của thực thể tại thời gian .
* là người đưa ra ý kiến.
* là thời điểm đưa ý kiến

Trong phân tích ý kiến được chia làm 4 hướng nghiên cứu chính cơ bản:

* Phân lớp chủ quan: xác định ý kiến là chủ quan hay khách quan
* Phân lớp cảm xúc: xác định ý kiến là tích cực hay tiêu cực
* Tóm tắt ý kiến: rút gọn nội dung ý kiến
* Khai thác ý kiến trên đặc trưng: tương tự phân lớp cảm xúc nhưng chi tiết hơn là xác định ý kiến tích cực hay tiêu cực trên đặc trưng nào.

Trong phân tích ý kiến ta có thể dễ dàng bắt gặp các từ như ý kiến (opinion), cảm nghĩ (sentiment), chủ quan (subjective) ở các tài liệu nghiên cứu. Những nghiên cứu này thường có tên gọi gắn liền với các cụm từ như khai thác ý kiến (opinion mining), phân tích cảm xúc (sentiment analysis) và phân tích chủ quan ( subjective analysis). Đây là những cơ sở quan trọng để tìm kiếm các tài liệu tham khảo trung cùng lĩnh vực.

Ngoài ra trong phân tích ý kiến còn có một số ý kiến mang tính chất riêng biệt như ý kiến so sánh.

Tuy nhiên các công trình nghiên cứu về phân tích ý kiến đa số được thực hiện trên các tập dữ liệu tiếng Anh, số nghiên cứu trên tập dữ liệu tiếng Việt vẫn còn hạn chế và cần được nghiên cứu đóng góp.

# Phân tích cảm xúc

Cảm xúc là suy nghĩ chủ quan của một con người về một khía cạnh nào đó. Theo nghiên cứu của Parrott [4], con người có sáu cảm xúc chính: tình yêu, niềm vui, bất ngờ, giận dữ, buồn bã và sợ hãi.

Phân tích cảm xúc (Sentiment analysis) là nhằm phát hiện ra thái độ mang tính lâu dài, màu sắc tình cảm, khuynh hướng niềm tin trong một vấn đề nào đó. Bài toán phân tích cảm xúc là bài toán dạng phân lớp cảm xức dựa trên văn bản ngôn ngữ tự nhiên. Đầu vào của bài toán là một câu hay một đoạn văn bản đầu ra là các giá trị xác suất của N lớp cảm xúc cần xác định.

Trong bài toán phân tích cảm xúc thường được phân thành các bài toán có độ khó như sau:

* Đơn giản: Phân tích cảm xúc thành 2 lớp là tích cực (positive) và tiêu cực (negative).
* Trung bình: Xếp hạng cảm xúc theo mức độ.
* Khó: Phát hiện mục tiêu nguồn gốc của cảm xúc hoặc các loại cảm xúc phức tạp.

Hiện tại đa số trong các nghiên cứu phân tích cảm xúc trên Tiếng Việt thường thực hiện bài toán ở cấp độ đơn giản là phân tích cảm xúc với 2 lớp cảm xúc tích cực hoặc tiêu cực. Trong nghiên cứu này cũng sẽ xây dựng bài toán phân tích cảm xúc ở mức độ đơn giản.



Hình 2.1 Mô hình xử lý Sentiment Analysis Vietnamese (SAV)

Hiện tại bài toán phân tích cảm xúc có thể được giải quyết dựa trên những phương pháp như:

* Theo phương pháp phân lớp không giám sát [5].
* Theo phương pháp phân lớp có giám sát [6]. Kỹ thuật chủ yếu dùng là Naïve Bayes hoặc SVM (support vector machines).
* Phân tích cảm xúc dựa trên khía cạnh. Một số kỹ thuật tiêu biểu của phương pháp này là dựa trên từ vựng [7].
* Phân loại cảm xác dựa trên chủ đề [8].

Nghiên cứu này sẽ tập trung nghiên cứu về các phương pháp phân lớp có giám sát phổ biến như: Naïve Bayes, SVM (support vector machines).

1. *Phương pháp phân lớp Naïve Bayes*

Naïve Bayes là một thuật toán máy học giám sát được sử dụng rộng rãi trong lĩnh vực máy học [9][10]. Ý tưởng cơ bản của cách tiếp cận này là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Với giả định này NB không sử dụng sự phụ thuộc của nhiều từ vào một chủ đề, không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề và do đó việc tính toán Naïve Bayes chạy nhanh hơn các phương pháp khác với độ phức tạp theo hàm số mũ.

Nhìn chung NB gán một tài liệu (biễu diễn bằng vector ) đến một lớp mà nó cực đại P( | ) theo luật Bayes như sau:

|)=

Trong đó:

* là xác suất ngẫu nhiên 1 tài liệu d có vector .
* là xác suất ngẫu nhiên một tài liệu thuộc lớp c.

Để tính được Navie Bayes đưa ra giả thuyết rằng tất cả đặc trưng trong là độc lập do đó ta có:

|)=

Ngoài ra còn có các phương pháp NB khác có thể kể ra như sau ML Naive Bayes, MAP Naive Bayes, Expected Naive Bayes, Bayesian Naive Bayes (Jason mô tả năm 2001). Naive Bayes là một công cụ rất hiệu quả trong một số trường hợp. Kết quả có thể sẽ bị giảm đi độ chính xác nếu dữ liệu huấn luyện hạn chế và các tham số dự đoán (như không gian đặc trưng) có chất lượng kém.

NB có ưu điểm là cài đặt đơn giản, tốc độ nhanh, dễ dàng cập nhật dữ liệu huấn luyện mới và có tính độc lập cao với tập huấn luyện, có thể sử dụng kết hợp nhiều tập huấn luyện khác nhau. Tuy nhiên nhằm mục đích cải thiện hiệu năng của NB các phương pháp như multiclass-boosting, ECOC (do Berger trình bày năm 1999 và Ghani mô tả lại năm 2000) có thể được dùng kết hợp.

1. *Phương pháp phân lớp SVM (support vector machines).*

Support vector Machine (SVM) là phương pháp tiếp cận phân lớp rất hiệu quả được Vapnik giới thiệu năm 1995 [11].

Ý tưởng của phương pháp này là 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 mặt phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp + (dương) và lớp – (âm). Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm được khoảng cách biên lớn nhất để tạo được kết quả phân loại tốt.



Hình Mô hình biễu diễn SVM.

Mô hình SVM[12] có thể được mô tả như sau:

Trong đó:

* là các vector đặc trưng.
* là các nhãn dán tương ứng.

Các siêu mặt phẳng ( trên hình) trong không gian đối tượng có phương trình là trong w là vector trọng số, b là độ dịch, không gian dữ liệu thuộc lớp âm thỏa mãn phương trình , không gian dữ liệu thuộc lớp dương thỏa mãn phương trình . Vì vậy bộ phân loại SVM được định nghĩa theo công thức:

Trong đó:

* sign(x) = +1 nếu z ≥ 0
* sign(x) = -1 nếu z < 0

Siêu phẳng là mặt phẳng đi qua các điểm thuộc lớp âm và có phương trình biểu diễn là , siêu phẳng là mặt phẳng đi qua các điểm thuộc lớp dương và có phương trình biểu diễn là .

Khoảng các từ 2 mặt phẳng và được gọi là biên (margin) và được tính theo công thức:

Trong đó: là độ dài của vector w.

Một mô hình SVM tối ưu là mô hình có giá trị margin đạt cực đại. Trong một số trường hợp để muốn có margin cao, ta chấp nhận việc một vài dữ liệu có thể không được chia chính xác (ví dụ như 1 dữ liệu + bị lọt sang vùng của -). Data này được gọi là nhiễu. Margin trong trường hợp này gọi là **Soft Margin**. **Hard Margin** ám chỉ việc tìm được margin mà không nhiễu (tất cả các dữ liệu đều thoả mãn sự phân lớp).

Với các bái toán thực tế việc tìm được **Hard Margin** nhiều khi là bất khả thi, vì thế việc chấp nhận sai lệch ở một mức độ chấp nhận được là vô cùng cần thiết.

Trong cài đặt SVM, người ta giới thiệu tham số C với quy ước:

* C =∞ không cho phép sai lệch, đồng nghĩa với Hard Margin.
* C lớn cho phép sai lệch nhỏ và giá trị margin nhỏ.
* C nhỏ cho phép sai lệch lớn và giá trị margin lớn

Có thể nói SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán là tìm được một không gian và siêu mặt phẳng quyết định sao cho sai số khi phân loại là thấp nhất, nghĩa là kết qủa phân loại sẽ cho kết quả tốt nhất.

Trong một số trường hợp SVM không thể phân chia dữ liệu bằng cách thông qua siêu mặt phẳng, SVM sẽ ánh xạ không gian ban đầu này sang một không gian khác với số chiều nhiều hơn sau đó tìm siêu mặt phẳng trong không gian này [13]. Kỷ thuật được sử dụng để thực hiện việc này là sử dụng hàm nhân (kernel function) thay cho tích có hướng của vector. Các hàm kernel phổ biến hiện nay trong SVM là :

* Linear
* Radial basis function
* Polyminal
* Sigmoid



Hình Minh họa các hàm kernel trong SVM

Trong nghiên cứu này sẽ tập trung tìm hiểu và áp dụng phương pháp phân lớp SVM Linear.

1. *Phương pháp K-Nearest Neighbor (k-NN)*

Đây là phương pháp truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trong nhận dạng mẫu hơn bốn thập kỷ qua (theo tài liệu của Dasarathy năm 1991). k-NN được đánh giá là một trong những phương pháp tốt nhất (áp dụng trên tập dữ liệu Reuters phiên bản 21450), được sử dụng từ những thời kỳ đầu của việc phân loại văn bản (được trình bày bởi Marsand năm 1992, Yang năm 1994, Iwayama năm 1995).

Ý tưởng chủ đạo của phương pháp này là khi cần phân loại một văn bản mới, thuật toán sẽ tính khoảng cách (khoảng cách Euclide, Cosine ...) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất (gọi là k “láng giềng”), sau đó dùng các khoảng cách này đánh trọng số cho tất cả chủ đề. Trọng số của một chủ đề chính là tổng tất cả khoảng cách ở trên của các văn bản trong k láng giềng có cùng chủ đề, chủ đề nào không xuất hiện trong k láng giềng sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo mức độ trọng số giảm dần và các chủ đề có trọng số cao sẽ được chọn là chủ đề của văn bản cần phân loại.

Xét chủ đề của văn bản khi đó trọng số của chủ đề sẽ được tính:

Trong đó:

* € {0,1} với y = 0 thì văn bản không thuộc về chủ đề , y = 1 thì văn bản thuộc chủ đề .
* độ giống nhau của văn bản và văn bản . Có thể sử dụng độ do0 cosine để tính .
* là ngưỡng phân loại của chủ đề được chọn ra từ tập huấn luyện.

KNN còn gọi là “Lazy learning method” vì tính đơn giản của nó, có nghĩa là quá trình training không quá phức tạp để hoàn thiện mô hình (tất cả các dữ liệu đào tạo có thể được sử dụng để kiểm tra mô hình KNN). Điều này làm cho việc xây dựng mô hình nhanh hơn nhưng giai đoạn thử nghiệm chậm hơn và tốn kém hơn về mặt thời gian và bộ nhớ lưu trữ, đặc biệt khi bộ dữ liệu lớn và phức tạp với nhiều biến khác nhau. Trong trường hợp xấu nhất, KNN cần thêm thời gian để quét tất cả các điểm dữ liệu và việc này sẽ cần nhiều không gian bộ nhớ hơn để lưu trữ dữ liệu.

Ngoài ra KNN không cần dựa trên các tham số khác nhau để tiến hành phân loại dữ liệu, không đưa ra bất kỳ kết luận cụ thể nào giữa biến đầu vào và biến mục tiêu, mà chỉ dựa trên khoảng cách giữa data point cần phân loại với data point đã phân loại trước đó. Đây là một đặc điểm cực kỳ hữu ích vì hầu hết dữ liệu trong thế giới thực tại không thực sự tuân theo bất kỳ giả định lý thuyết nào ví dụ như phân phối chuẩn trong thống kê.

Bước khó khăn nhất của thuật toán KNN và cũng là bước đau đầu nhất đó chính là chọn K là bao nhiêu. K càng lớn độ chính xác của thuật toán sẽ càng được cải thiện.

Xét ví dụ sau giả sử một bệnh viện tiến hành phân loại thuốc chỉ định cho những bệnh nhân mới dựa trên độ tuổi (Age) và tỷ lệ Na/K trong máu.



Bên trên là độ thị Scatter Plot, trục hoành là độ tuổi, trục tung là tỷ lệ Na/K, mỗi điểm trên đồ thị là một bệnh nhân tương ứng với tỷ lệ Na/K, và độ tuổi cho trước. Màu sắc khác nhau thể hiện cho loại thuốc chỉ định. Màu xanh nhạt là loại thuốc M, màu xanh trung bình là loại thuốc N, màu xanh đậm là loại thuốc P

Giả sử bệnh viện tiếp nhận các bệnh nhân mới (ví dụ bệnh nhân A và B) và cần tiến hành phân loại thuốc cho họ. Đồ thị tiếp theo dưới đây chứa các bệnh nhân mới chưa được phân loại, dựa vào độ tuổi, và tỷ lệ Na/K chúng ta xác định được những vùng trên đồ thị sẽ là nơi chứa các data point của các bệnh nhân mới này. Nhiệm vụ là xác định loại thuốc thích hợp cho bệnh nhân A, B trên cơ sở là xác định khoảng cách giữa điểm dữ liệu của A, và B (chưa được phân loại thuốc) và điểm dữ liệu đã phân loại (bệnh nhân cũ trước đây đã được phân loại thuốc), khoảng cách gần nhất thì khả năng loại thuốc được phân loại sẽ tương đương nhau giữa 2 bệnh nhân.

Trước tiên chúng ta xét bệnh nhân A giả sử có độ tuổi là 40 và tỷ lệ Na/K gần 29, thì thấy rằng điểm dữ liệu của bệnh nhân này nằm trong vùng chứa có các điểm dữ liệu màu xanh nhạt tức nằm chung vùng với các bệnh nhân trước đây được phân loại thuốc là M. Do đó bệnh nhân mới số 1 sẽ được phân loại thuốc là M. Ở đây chúng ta không cần đặt giá trị K để tìm ra các điểm gần nhất do xung quanh của điểm dữ liệu bệnh nhân A toàn là các điểm màu xanh nhạt. (Hình đồ thị dưới đây được điều chỉnh màu sắc để các bạn nhìn rõ các điểm)



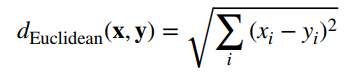
Xét tiếp bệnh nhân B, lưu ý hình đã được chúng tôi cân chỉnh lại màu sắc để hiển thị rõ màu sắc khác nhau giữa các điểm giúp các bạn dễ phân biệt. Giả sử zoom lại gần vùng chứa điểm dữ liệu của bệnh nhân B chúng ta có hình dưới đây.



Nếu chúng ta lấy K = 1 tức chỉ xét 1 điểm gần nhất so với điểm dữ liệu của bệnh nhân B, thì điểm dữ liệu bệnh nhân B gần nhất với điểm màu xanh đậm nhất, ứng với bệnh nhân B sẽ được phân loại thuốc là P. Nếu chúng ta lấy K = 2 tức xét 2 điểm gần nhất, thì điểm dữ liệu B sẽ gần với 1 điểm xanh đậm và 1 điểm màu xanh trung bình, tức là bệnh nhân B có thể được phân loại thuốc là P hoặc là N. Do đó chúng ta chưa tìm ra đâu là loại thuốc thích hợp nhất cho B, vậy K = 2 không phải là giá trị K cần xét. Tiếp đến chúng ta lấy K = 3, thì trên đồ thị chúng ta thấy có 2 điểm màu xạnh trung bình nhiều hơn so với 1 điểm màu xanh đậm là gần nhất với điểm dữ liệu bệnh nhân B. Vậy với K = 3, bệnh nhân B sẽ được phân loại thuốc là N khi điểm dữ liệu bệnh nhân B gần với nhiều điểm dữ liệu màu xanh trung bình hơn.

Phương pháp trên gọi là Voting, tức tìm ra những điểm dữ liệu phổ biến xuất hiện gần nhất với điểm dữ liệu cần phân loại (trong thống kê cũng có thể gọi là tính Mode). Trở lại ví dụ trên thì với K = 3, số vote cho điểm dữ liệu màu xanh trung bình là 2, còn điểm dữ liệu màu xanh đậm là 1, vậy 2 > 1 nên bệnh nhân B sẽ được phân loại thuốc là P, độ tin cậy Confidence = 2/3 = 66.7%.

Lưu ý quan trọng, chúng ta thường tìm K trước rồi mới khoanh vùng cho điểm dữ liệu chưa phân loại dựa trên việc tính toán các khoảng cách giữa nó so với các điểm dữ liệu đã phân loại. Trong ví dụ này việc tính khoảng cách giữa các điểm dữ liệu dựa trên phương pháp Euclidean:



Ví dụ bệnh nhân mới D có tuổi là 20, và tỷ lệ Na/K là 12, bệnh nhân cũ E có tuổi là 30, tỷ lệ Na/K là 8. Vậy khoảng cách

Trong trường hợp sau khi tính toán khoảng cách của 3 điểm gần nhất với bệnh nhân D cho ra 3 điểm dữ liệu màu hoàn toàn khác nhau vậy ta sẽ không xác định được loại thuốc nào nào cho bệnh nhân D. Ở trường hợp này ta sẽ xem xét gán trọng số vào các thuộc tính theo mức độ ảnh hưởng của thuộc tính tới thuộc tính phân loại, ở ví dụ này tỷ lệ Na/K quan trọng hơn độ tuổi trong việc phân loại do đó gán trọng số là 3 và độ tuổi là 1. Tính lại khoảng cách giữa bệnh nhân mới D và bệnh nhân cũ E đã được phân loại thuốc trước đó chúng ta có:

Ưu điểm của KNN:

* Độ phức tạp tính toán của quá trình training là bằng 0.
* Việc dự đoán kết quả của dữ liệu mới rất đơn giản.

Nhược điểm của KNN:

* KNN rất nhạy cảm với nhiễu khi K nhỏ.
* Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

1. *Phương pháp Phương pháp Linear Least Square Fit (LLSF)*

LLSF là một cách tiếp cận ánh xạ được phát triển bởi Yang và Chute vào năm 1992. Ban đầu LLSF được thử nghiệm trong lĩnh vực xác định từ đồng nghĩa sau đó sử dụng trong phân loại vào năm 1994. Các thử nghiệm cho thấy hiệu suất phân loại của LLSF có thể ngang bằng với phương pháp K-NN kinh điển.

Ý tưởng của LLSF là sử dụng phương pháp hồi quy để học từ tập huấn luyện và các chủ đề có sẵn.

Tập huấn luyện được biểu diễn dưới dạng một cặp vector đầu vào và đầu ra như sau:

* Vector đầu vào là một văn bản bao gồm các từ và trọng số.
* Vector đầu ra gồm các chủ đề cùng với trọng số nhị phân của văn bản ứng với vector đầu vào.
* Giải phương trình các cặp vector đầu vào, đầu ra chúng ta sẽ thu được ma trận đồng hiện của hệ số hồi quy của từ và chủ đề.

Phương pháp này sử dụng công thức:

Trong đó :

* A, B là ma trận đại diện tập dữ liệu huấn luyện (các cột trong ma trận tương ứng là các vector đầu vào và đầu ra).
* là ma trận kết quả chỉ ra một ánh xạ từ một văn bản bất kỳ vào vector của chủ đề đã gán trọng số.

Nhờ vào việc sắp xếp trọng số của các chủ đề, chúng ta được một danh sách chủ đề có thể gán cho văn bản cần phân loại. Nhờ đặt ngưỡng lên trọng số của các chủ đề mà ta tìm được chủ đề thích hợp cho văn bản đầu vào. Hệ thống tự động học các ngưỡng tối ưu cho từng chủ đề, giống với KNN. Mặc dù LLSF và KNN khác nhau về mặt thống kê, nhưng chúng ta vẫn tìm thấy điểm chung trong cách làm của hai phương pháp này là quá trình học ngưỡng tối ưu.



# Phân lớp câu chủ quan

Câu chủ quan là một câu thể hiện về cảm xúc hoặc ý kiến rõ ràng của một cá nhân. Phải phân biệt rõ 1 câu là thể hiện ý kiến khách quan hay chủ quan.

Ở kỹ thuật này ta dùng đa số bằng phương pháp Navie Bayes [9]. Với kỷ thuật trên ý kiến không được đưa về bộ 5 thành phần như ở phân tích ý kiến. Ngoài ra còn một số phương pháp khác như:

* Phân loại câu chủ quan sử dụng mẫu [10].
* Câu chủ quan và phân lớp cảm xúc [11].

# Các mô hình biểu diễn văn bản

### *Mô hình logic*

Theo mô hình này, các từ có nghĩa trong văn bản sẽ được đánh chỉ số và nội dung văn bản được quản lý theo các chỉ số Index đó. Mỗi văn bản được đánh chỉ số theo quy tắc liệt kê các từ có nghĩa trong các văn bản với vị trí xuất hiện của nó trong văn bản. Từ có nghĩa là từ mang thông tin chính về các văn bản lưu trữ, khi nhìn vào nó, người ta có thể biết chủ đề của văn bản cần biễu diễn.

Khi đó chúng ta tiến hành Index các văn bản đưa vào theo danh sách các từ khóa nói trên. Với mỗi từ khóa người ta sẽ đánh số thứ tự vị trí xuất hiện của nó và lưu lại chỉ số đó cùng với mã văn bản chứa nó. Cách biểu diễn này cũng được các máy tìm kiếm ưa dùng.

Ví dụ: Có 2 văn bản với mã tương ứng là VB1, VB2:

VB1 là: “Đại hội chi bộ thành công”

VB2 là: “Chi bộ hoàn thành nhiệm vụ”

Khi đó, ta có cách biểu diễn như sau:

Bảng 2.1: Biểu diễn văn bản trong mô hình Logic

|  |  |
| --- | --- |
| Từ mục | Mã VB\_Vị trí xuất hiện |
|  |  |
| Đại | VB1(1) |
|  |  |
| Hội | VB1(2) |
|  |  |
| Chi | VB1(3), VB2(1) |
|  |  |
| Bộ | VB1(4), VB2(2) |
|  |  |
| Thành | VB1(5),VB2(4) |
|  |  |
|  | 7 |

Công VB1(6)

Hoàn VB2(3)

Nhiệm VB2(5)

Vụ VB2(6)

**Ưu điểm, nhược điểm của mô hình logic:**

**+ Ưu điểm**

Việc tìm kiếm trở nên nhanh chóng và đơn giản. Cần tìm kiếm từ “computer”. Hệ thống sẽ duyệt trên bảng Index để trỏ đến chỉ số Index tương ứng nếu từ “computer” tồn tại trên hệ thống. Việc tìm kiếm này khá nhanh và đơn giản khi trước đó ta đã sắp xếp bảng Index theo vẫn chữ cái. Phép tìm kiếm trên có độ phức tạp cấp (*n* log2 *n*) , với n là số từ trong bảng Index. Tương ứng với chỉ số index trên sẽ cho ta biết các tài liệu chứa từ khóa tìm kiếm. Như vậy, việc tìm kiếm liên quan đến k từ thì các phép toán cần thực hiện

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| là k\*n\* | log | 2 | *n* | (với n là số từ trong bảng Index). |  |
|  |  |  |

**+ Nhược điểm**

Với phương pháp này đòi hỏi người sử dụng phải có kinh nghiệm và chuyên môn trong lĩnh vực tìm kiếm vì câu hỏi đưa vào dưới dạng Logic nên kết quả cũng có giá trị Logic (Boolean). Một số tài liệu sẽ được trả lại khi thỏa mãn mọi điều kiện đưa vào. Như vậy muốn tìm được tài liệu theo nội dung thì phải biết đích xác về tài liệu.

Việc Index các tài liệu rất phức tạp và làm tốn nhiều thời gian, đồng thời cũng tốn không gian để lưu trữ các bảng Index.

Các tài liệu tìm được không được sắp xếp theo độ chính xác của chúng. Các bảng Index không linh hoạt vì khi các từ vựng thay đổi (thêm, sửa, xóa, ...) dẫn tới chỉ số Index cũng phải thay đổi theo.

### *Mô hình phân tích cú pháp*

Trong mô hình này, mỗi văn bản đều phải được phân tích cú pháp và trả lại thông tin chi tiết về chủ đề của văn bản đó. Sau đó, người ta tiến hành Index các chủ đề của từng văn bản. Cách Index trên chủ đề cũng giống như Index trên văn bản nhưng chỉ Index trên các từ xuất hiện trong chủ đề.

Các văn bản được quản lý thông qua các chủ đề này để có thể tìm kiếm được khi có yêu cầu, câu hỏi tìm kiếm sẽ dựa trên các chủ đề trên.

**Một số ưu điểm, nhược điểm của phương pháp này:**

**+ Ưu điểm**

Tìm kiếm theo phương pháp này khá hiệu quả và đơn giản, do tìm kiếm nhanh và chính xác.

Đối với những ngôn ngữ đơn giản về mặt ngữ pháp thì việc phân tích trên có thể đạt được mức độ chính xác cao và chấp nhận được.

**+ Nhược điểm**

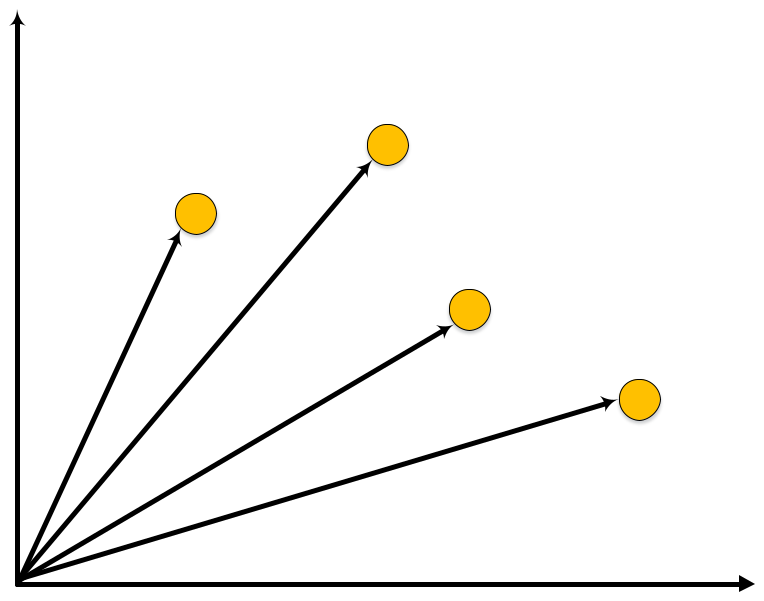
Chất lượng của hệ thống theo phương pháp này hoàn toàn phụ thược vào chất lượng của hệ thống phân tích cú pháp và đoán nhận nội dung tài liệu. Trên thực tế, việc xây dựng hệ thống này rất phức tạp, phụ thuộc vào đặc điểm của từng ngôn ngữ và đa số chưa đạt đến độ chính xác cao.

### *Mô hình không gian vector*

Cách biểu diễn văn bản thông dụng nhất là thông qua vector biểu diễn theo mô hình không gian vector (Vector Space Model). Đây là một cách biểu diễn tương đối đơn giản và hiệu quả.

Theo mô hình này, mỗi văn bản được biểu diễn thành một vector. Mỗi thành phần của vector là một từ khóa riêng biệt trong tập văn bản gốc và được gán một giá trị là hàm f chỉ mật độ xuất hiện của từ khóa trong văn bản.

9



Văn bản 2

Văn bản 1

Văn bản 3

Văn bản 4

Hình 2.1 Biểu diễn vector văn bản trong không gian 2 chiều

Giả sử ta có một văn bản và nó được biểu diễn bởi vector *V* (*v*1 , *v*2 ,..., *vn* ). Trong đó,

vi là số lần xuất hiện của từ khóa thứ i trong văn bản. Ta xét 2 văn bản sau:

VB1: Đại hội chi bộ

VB2: Đại hội đã thành công

Sau khi qua bước tiền xử lý văn bản, ta biểu diễn như sau:

Bảng 2.2: Biểu diễn văn bản mô hình Vector

|  |  |  |
| --- | --- | --- |
| Từ | Vector\_VB1 | Vector\_VB2 |
|  |  |  |
| Đại | 1 | 1 |
|  |  |  |
| Hội | 1 | 1 |
|  |  |  |
| Chi | 1 | 0 |
|  |  |  |
| Bộ | 1 | 0 |
|  |  |  |
| Thành | 0 | 1 |
|  |  |  |
| Công | 0 | 1 |
|  |  |  |
|  | 10 |  |

Trong các cơ sở dữ liệu văn bản, mô hình vector là mô hình biểu diễn văn bản được sử dụng phổ biến nhất hiện nay. Mối quan hệ giữa các văn bản được thực hiện thông qua việc tính toán trên các vector biểu diễn vì vậy được thi hành khá hiệu quả.

### *Mô hình Boolean*

Một mô hình biểu diễn vector với hàm f cho ra giá trị rời rạc với duy nhất hai giá trị đúng và sai (true và false, hoặc 0 và 1) gọi là mô hình Boolean. Hàm f tương ứng với từ khóa ti sẽ cho ra giá trị đúng nếu và chỉ nếu từ khóa ti xuất hiện trong văn bản đó.

Mô hình Boolean được xác định như sau:

Giả sử có một cơ sở dữ liệu gồm ***m*** văn bản, ***D = {d1, d2, ..., dm}***. Mỗi văn bản được biểu diễn dưới dạng một vector gồm ***n*** từ khóa ***T = {t1, t2, ..., tn}***. Gọi ***W = {Wij}*** là ma trận trọng số, trong đó ***Wij*** là giá trị trọng số của từ khóa ***ti*** trong văn bản ***dj***.

|  |  |  |  |
| --- | --- | --- | --- |
| ij | = |  |  |
| W |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 *neáu t* | *coùmaëttrongd* | *j* |  |
| *i* |  |  |
| 0 *neáungöôïc laïi* | |  |  |

Trở lại với 2 văn bản trên, áp dụng mô hình Boolean ta có biểu diễn như sau:

Bảng 2.3: Biểu diễn văn bản mô hình Boolean

|  |  |  |
| --- | --- | --- |
| Từ | Vector\_VB1 | Vector\_VB2 |
|  |  |  |
| Đại | 1 | 1 |
|  |  |  |
| Hội | 1 | 1 |
|  |  |  |
| Chi | 1 | 0 |
|  |  |  |
| Bộ | 1 | 0 |
|  |  |  |
| Thành | 0 | 1 |
|  |  |  |
| Công | 0 | 1 |
|  |  |  |

### *Mô hình tần số từ khóa (TF – Term Frequency)*

### *Mô hình nghịch đảo tần số văn bản (IDF – Inverse Document Frequency)*

### *Mô hình TF - IDF*

# 

# Phân lớp câu chủ quan

**CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN**

**CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ**

**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

# TÀI LIỆU THAM KHẢO

1. B. Jindal & B. Liu, Mining Comparative Sentences and Relations, American Association for Artificial Intelligence, Pages 1331-1336, 2006.
2. M. Hu & B. Liu, Mining and summarizing customer reviews, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 168-177, 2004.
3. B. Liu , Sentiment analysis and subjectivity, Handbook of Natural Language Processing, 2010.
4. J. Parrott & A. Bourne & R. Akien & J. Irvine, Self-Optimizing Continuous Reactions in Supercritical Carbon Dioxide, [Angewandte Chemie International Edition](https://www.researchgate.net/journal/1521-3773_Angewandte_Chemie_International_Edition), Pages 3788-3792, 2010.
5. B. Pang & L. Lee & S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP, Page 79-86, 2002.
6. D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Page 417-424, 2002.
7. G. Qiu & B. Liu & J. Bu & C. Chen, Opinion word expansion and target extraction through double Propagation, Journal Computational Linguistics, Page 9-27, 2011.
8. X. Ding & B. Liu & S. Yu, A holistic lexicon approach to opinion mining, Proceedings of the 2008 International Conference on Web Search and Data Mining, Page 231-240, 2008.
9. Tang H., Tan S., and Cheng X. (2009), “A survey on sentiment detection of reviews”, *Expert Systems with Applications*, Vol. 36, No. 7, pages 10760-10773.
10. Stanford University (2019). Text Classification and Naïve Bayes [online], viewed 12 March 2019, from:< <https://web.stanford.edu/class/cs124/lec/naivebayes.pdf>>
11. V.N. Vapnik, “The Nature of Statistical Learning Theory,” Springer, New York, 1995.
12. Y.Gao, S.Sun , An Empirical Evaluation of Linear and Nonlinear Kernels for Text Classification Using Support Vector Machines, Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010
13. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, 2005

J. Wiebe & F. Bruce & P. O’Hara, Development and use of a gold standard data set for subjectivity classifications, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Page 246-253, 1999.