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
(Họ tên và chữ ký)

KHOA QUẢN LÝ CHUYÊN NGÀNH
(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

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
	TF	Term Frequency	
	IDF	Inverse Document Frequency	
	CGs	Conceptual Graphs	
	BOW	Bag of word	Túi từ

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

CHƯƠNG 1: GIỚI THIỆU

1.1 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.

1.2 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.

1.3 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.

1.4 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.

1.5 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.

1.6 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

2.1 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]:

$$e_{j,}a_{jk}, so_{ijkl}, h_i, t_l$$

Trong đó:

- e_j là một thực thể đích.
- a_{jk} là một khía cạnh/tính năng của thực thể e_j.
- so_{ijkl} là giá trị cảm xúc của ý kiến từ người giữ ý kiến h_i về tính năng a_{jk} của thực thể e_j tại thời gian t_l.
- h_i là người đưa ra ý kiến.
- t_l 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.

Trong bài toán phân tích ý kiến bao gồm nhiều bài toán nhỏ như: phân lốp chủ quan và khách quan (subjectivity classification), phân lớp ý kiến trái chiều (sentiment polarity classification), phát hiện ý kiến rác (spam opinion detection), tóm tắt và tổng hợp quan điểm (opinion summarization),…

Quan điểm trong phân tích thường được chia làm hai loại: tích cực (positive) và tiêu cực (negative). Tuy nhiên ngoài hai trạng thái này một ý kiến còn có thể ở trạng thái trung lập (neutral).

Bài toán phân tích ý kiến thường được tiếp cận và giải quyết ở ba mức độ:

- Mức độ văn bản, tài liệu (Document level): ở mức độ này, bài toán cần phân loại xem một văn bản hay tài liệu thể hiện ý kiến tiêu cực hay tích cực. Ví dụ như một bài viết phân tích, đánh giá về chất lượng giảng dạy mỗi học kỳ tại trường Đại học Công nghệ TP.HCM nhận định chủ yếu là tốt hay không tốt, tích cực hay tiêu cực. Mức độ này được thực hiện với giả sử rằng tài liệu chỉ đưa ra các quan điểm, ý kiến về một thực thể duy nhất chứ không có sự so sánh giữa các thực thể khác nhau.
- Mức độ câu (Sentence level): các phương pháp được áp dụng cho mức độ tài liệu cũng có thể được áp dụng ở mức độ câu. Trong trường hợp đơn giản, các câu chỉ chứa một ý kiến, quan điểm về một thực thể. Trong các trường hợp phức tạp hơn, một câu có thể có nhiều quan điểm, đánh giá về các khía cạnh khác nhau của một đối tượng hoặc thậm chí có thể có sự thay đổi về quan điểm trong cùng một câu (polarity shifting). Mức độ phân tích quan điểm cho câu rất gần với bài toán phân lớp chủ quan và khách quan, trong đó chúng ta cần phân loại xem một câu đã cho là chủ quan (có quan điểm, ý kiến riêng) hay khách quan (câu chỉ đưa ra thông tin). Tuy nhiên, các câu khách quan cũng có thể từ đó suy ra quan điểm. Ví dụ như câu: Chiếc xe tôi mua tháng trước và cái kính chắn gió đã rơi ra. Trong câu nói này mệnh đề đều là sự việc khách quan nhưng trong thực tế nhưng từ đó có thể suy luận ra ý kiến chê bai chất lượng xe của nhà sản xuất.

- Mức độ khía cạnh (Aspect level): nếu với hai mức độ nêu trên, vấn đề được tiếp cận theo hướng kiến trúc của văn bản, ngôn ngữ (câu, đoạn, tài liệu, cú pháp), thì ở mức độ khía cạnh, bài toán tập trung vào chính quan điểm, ý kiến được đưa ra, phân tích ở mức độ sâu hơn, đó là phân tích xem ý kiến tiêu cực hay tích cực của là về chủ đề, đối tượng nào. Ví dụ: Giảng viên môn Tiếng Anh của tôi dạy phần nghe rất khó hiểu.

Phân tích ý kiến tuy đang là xu hướng hiện nay nhưng các công trình nghiên cứu 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 mở rộng hơn nữa.

2.2 Phân tích ý kiến

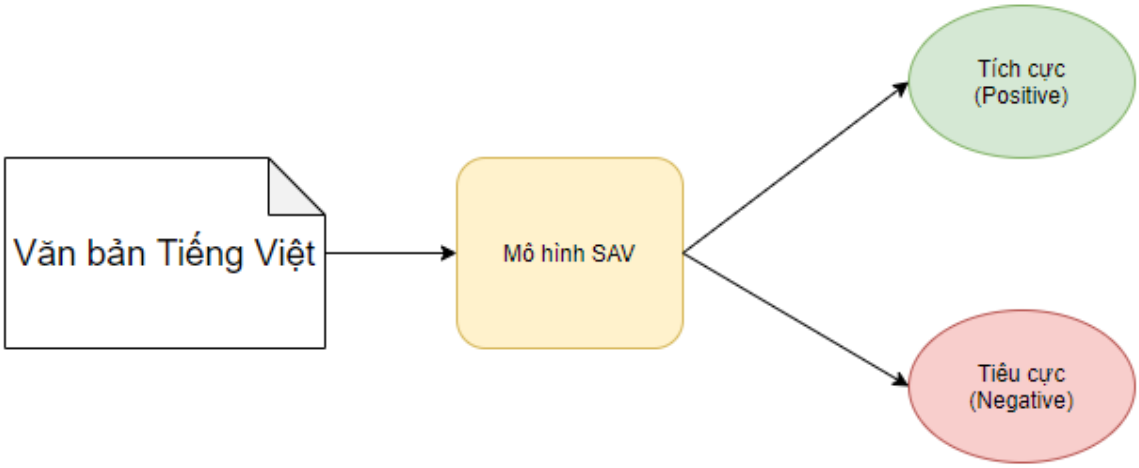
Cảm xúc là suy nghĩ chủ quan của một con người về một khía cạnh nào đó. Theo nghiên cứu của Parrott [4], con người có sáu cảm xúc chính: tình yêu, niềm vui, bất ngờ, giận dữ, buồn bã và sợ hãi.

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 xúc 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).

2.2.1 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 d_j (biểu diễn bằng vector d_j^*) đến một lớp c_i mà nó cực đại P($c_i \mid d_j^*$) theo luật Bayes như sau:

$$P(c_i|d_j^*)= \frac{P(c_i)P(d_j^* \mid c_i)}{P(d_j^*)}$$

Trong đó:

- $P(d_j^*)$ là xác suất ngẫu nhiên 1 tài liệu d có vector d_j^*.
- $P(c_i)$ là xác suất ngẫu nhiên một tài liệu thuộc lớp c.

Để tính được $P(d_j^* \mid c_i)$ Navie Bayes đưa ra giả thuyết rằng tất cả đặc trưng trong d_j^* là độc lập do đó ta có:

$$P(c_i|d_j^*)= \frac{P(c_i)P(\prod_{i=1}^{m} d_j^* \mid c_i)}{P(d_j^*)}$$

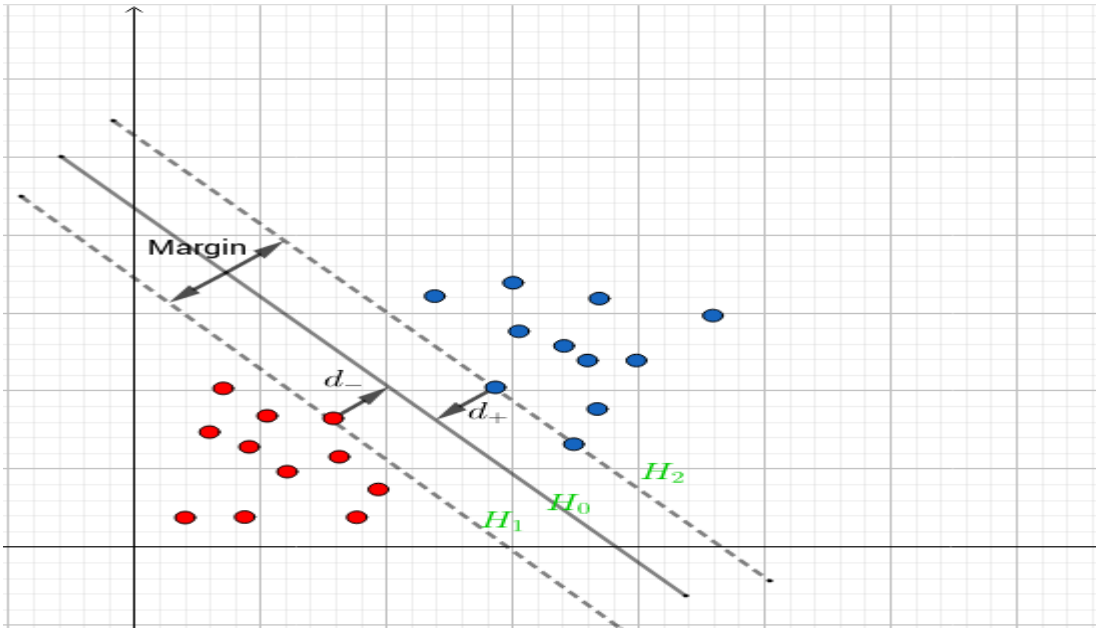
Ngoài ra còn có các phương pháp NB khác có thể kể ra như sau ML Naive Bayes, MAP Naive Bayes, Expected Naive Bayes, Bayesian Naive Bayes (Jason mô tả năm 2001). Naive Bayes là một công cụ rất hiệu quả trong một số trường hợp. Kết quả có thể sẽ bị giảm đi độ chính xác nếu dữ liệu huấn luyện hạn chế và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

2.2.2 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 2.2 Mô hình biểu diễn SVM.

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

$$\{(x_i, y_i), i = 1,2 \dots, i\}$$

Trong đó:

- x_i là các vector đặc trưng.
- y_i là các nhãn dán tương ứng.

Các siêu mặt phẳng (H_0 trên hình) trong không gian đối tượng có phương trình là $w^T x + b = 0$ trong w là vector trọng số, b là độ dịch, không gian dữ liệu thuộc lớp âm

thỏa mãn phương trình $w^T x + b \leq -1$, không gian dữ liệu thuộc lớp dương thỏa mãn phương trình $w^T x + b \geq 1$. Vì vậy bộ phân loại SVM được định nghĩa theo công thức:

$$f(x) = sign(w^T x + b)$$

Trong đó:

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

Siêu phẳng H_1 là mặt phẳng đi qua các điểm thuộc lớp âm và có phương trình biểu diễn là $w^T x + b = -1$, siêu phẳng H_2 là mặt phẳng đi qua các điểm thuộc lớp dương và có phương trình biểu diễn là $+b = 1$.

Khoảng các từ 2 mặt phẳng H_1và H_2 được gọi là biên (margin) và được tính theo công thức:

$$margin = \frac{2}{\|w\|^2} = \frac{2}{(w^T x)}$$

trong đó: $\|w\|$ 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 bài 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.

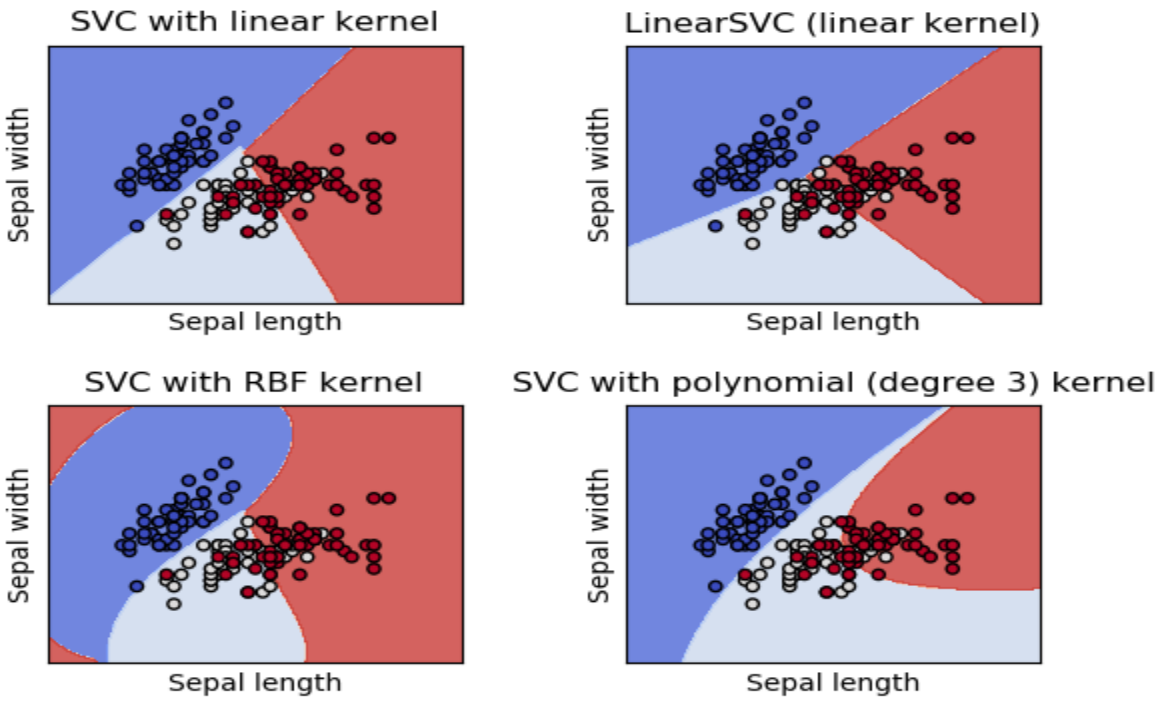
Trong cài đặt SVM, người ta giới thiệu tham số C với quy ước:

- 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 [12]. 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 2.3 Minh họa các hàm kernel trong SVM.

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

2.2.3 Phương pháp K-Nearest Neighbor (KNN)

Phương pháp K-Nearest Neighbor [13] 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). KNN 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 c_j của văn bản $\vec{x}$ khi đó trọng số của chủ đề sẽ được tính:

$$x(\vec{x}, c_j) = \sum sim(\vec{x}, \vec{d}_i).y(\vec{x}, c_j) - b_j$$

Trong đó:

- $y(\vec{x}, c_j)$ € {0,1} với y = 0 thì văn bản $\vec{x}$ không thuộc về chủ đề c_j, y = 1 thì văn bản $\vec{x}$ thuộc chủ đề c_j.
- $sim(\vec{x}, \vec{d}_i)$ độ giống nhau của văn bản $\vec{x}$ và văn bản $\vec{d}_i$. Có thể sử dụng độ đo cosine để tính $sim(\vec{x}, \vec{d}_i)$ như sau:

$$sim(\vec{x}, \vec{d}_i) = cos(\vec{x}, \vec{d}_i)$$

- b_j là ngưỡng phân loại của chủ đề c_j được chọn ra từ tập huấn luyện.

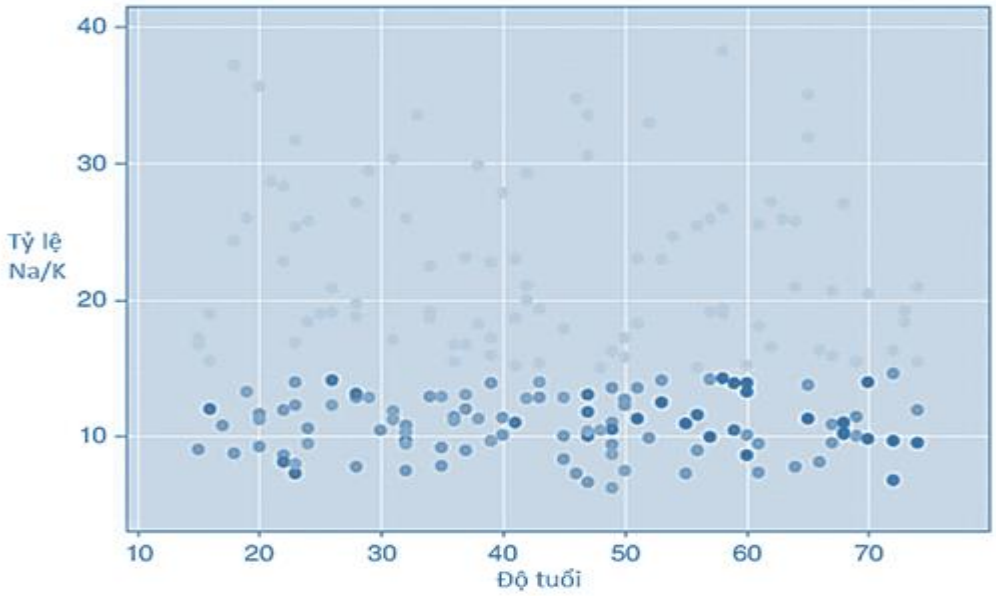
KNN còn gọi là "Lazy learning method" vì tính đơn giản của nó, có nghĩa là quá trình training không quá phức tạp để hoàn thiện mô hình (tất cả các dữ liệu đào tạo có thể được

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 [13].



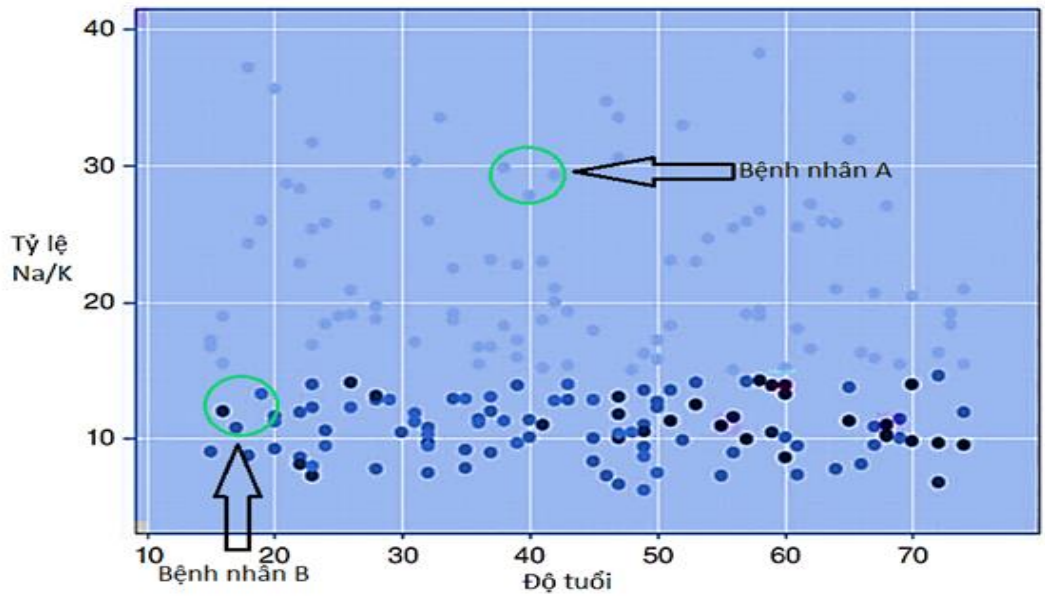
Hình 2.4 Biểu đồ huấn luyện bài toán chỉ định thuốc.

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.

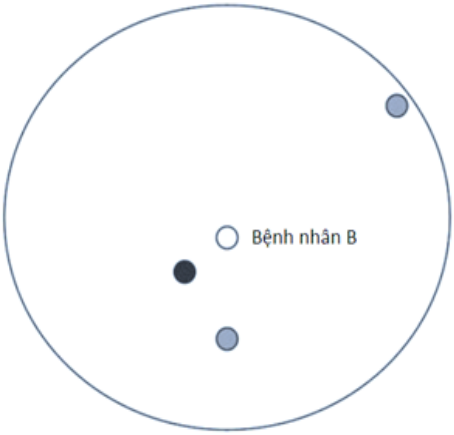
Giả sử bệnh viện tiếp nhận các bệnh nhân mới (ví dụ bệnh nhân A và B) và cần tiến hành phân loại thuốc cho họ. Đồ thị tiếp theo dưới đây chứa các bệnh nhân mới chưa được phân loại, dựa vào độ tuổi, và tỷ lệ Na/K chúng ta xác định được những vùng trên đồ thị sẽ là nơi chứa các data point của các bệnh 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).



Hình 2.5 Biểu đồ huấn luyện bài toán chỉ định thuốc 1.

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.



Hình 2.6 Biểu đồ chi tiết KNN của bệnh nhân B.

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:

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

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 sẽ là:

$$d_{Euclidean} = \sqrt{(20 - 30)^2 + (8 - 12)^2} = 10{,}77$$

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:

$$d_{Euclideean} = \sqrt{(20 - 30)^2 + 3 * (8 - 12)^2} = 12{,}1$$

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.

2.2.4 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 KNN 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: $F_{LS} = argmin = \|FA - B\|^2$

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).
- F_{LS} 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.

2.2.5 Phương pháp Entropy cực đại

Phương pháp Entropy cực đại là một kỹ thuật dùng để ước lượng xác suất các phân phối từ dữ liệu [14]. Tư tưởng chủ đạo của nguyên lý Entropy cực đại là "mô hình phân phối đối với mỗi tập dữ liệu và tập các ràng buộc đi cùng phải đạt được độ cân bằng / đều nhất có thể". Tập dữ liệu huấn luyện được sử dụng để tìm ra các ràng buộc cho mô hình, đó là cơ sở để ước lượng phân phối cho từng lớp cụ thể. Những ràng buộc này được thể hiện bởi các giá trị ước lượng được của các đặc trưng. Từ các ràng buộc sinh ra bởi tập dữ liệu này, mô hình sẽ tiến hành tính toán để có được một phân phối cho Entropy cực đại.

Theo nghiên cứu [15] thì các hàm đặc trưng $f(x, y)$ là một hàm nhị phân với 2 tham số: y ∈ tập các lớp cần phân loại và x ∈ tập ngữ cảnh:

$$f = \varepsilon \rightarrow \{0,1\}$$

Giá trị kỳ vọng của f có phân phối xác suất quan sát được $\tilde{p}(x, y)$là:

$$E_{\tilde{p}} f_i = \sum_{a,b} \tilde{p}(x, y) f(x, y)$$

Mọi tri thức quan sát được từ tập mẫu đều có thể được biểu diễn dưới dạng giá trị kỳ vọng của một hàm đặc trưng $f$phù hợp.

Với k đặc trưng các ràng buộc được biểu diễn dưới điều kiện:

$$E_p f_i = E_{\tilde{p}} f_i$$

với $0 \leq i \leq k$, $\tilde{p}$ là xác suất quan sát được của tập huấn luyện.

Việc lựa chọn các hàm đặc trưng tùy thuộc vào từng bài toán khác nhau và việc lựa chọn đặc trưng này ảnh hưởng đến chất lượng bộ phân lớp.

2.3 Phân lớp câu chủ quan

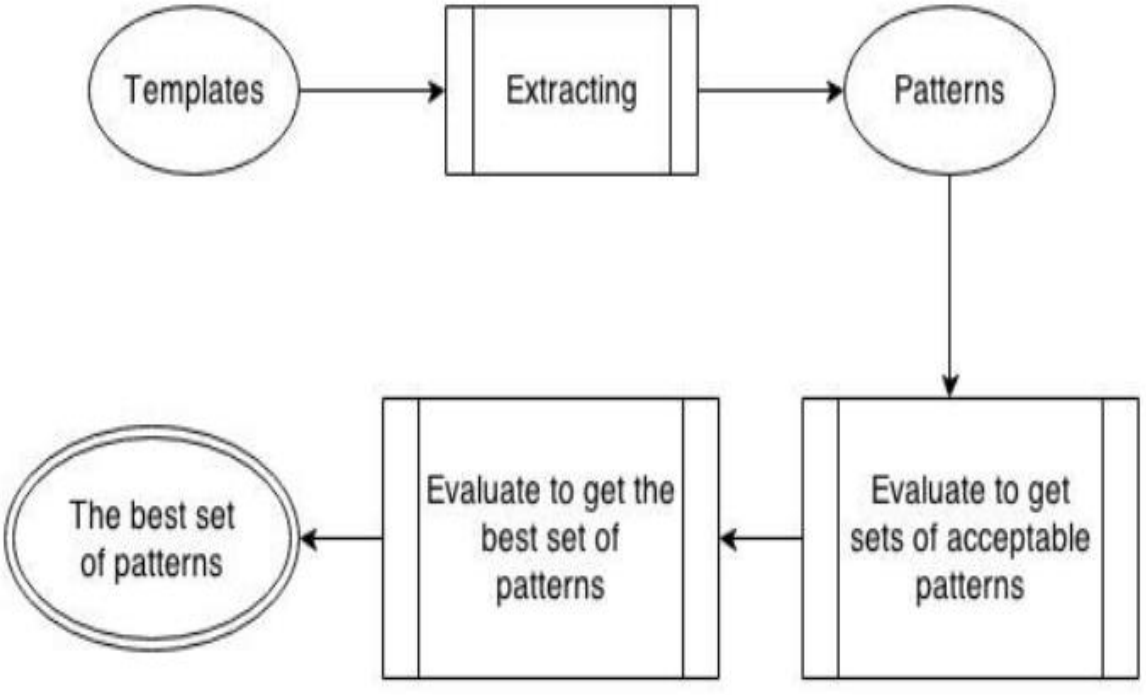
Câu chủ quan là một câu thể hiện về cảm xúc hoặc ý kiến rõ ràng của một cá nhân. Phân lớp câu chủ quan là xác định câu thuộc 1 trong hay lớp chủ quan hoặc khách quan (theo Wiebe vào 1999) [5]. Một câu khách quan thường diễn đạt đưa ra một số thông tin thực tế, trong khi câu chủ quan thường đưa ra những quan điểm và ý kiến cá nhân.

- Câu chủ quan: Tôi thích chiếc điện thoại iphone này.
- Câu khách quan: Chiếc iphone này có màu đỏ.

Trong thực tế câu chủ quan có thể diễn đạt nhiều loại thông tin khác nhau như: ý kiến, đánh giá, cảm xúc, niềm tin, suy đoán, cáo buộc, lập trường, v.v. (Quirk et al., 1985; Wiebe, Bruce and O'Hara, 1999). Trước đây trong một số nghiên cứu người ta đã xem việc phân lớp câu chủ quan là một vấn đề độc lập với việc phân loại tình cảm nhưng gần đây những nghiên cứu đã xem phân lớp câu chủ quan là một bước đầu tiên của việc phân tích cảm xúc bằng cách sử dụng nó để loại bỏ các câu khách quan được cho là không có ý kiến.

Đa số cách tiếp cận giải quyết bài toán phân lớp câu chủ quan là phân loại dựa trên học có giám sát đòi hỏi dữ liệu huấn luyện phải được gán nhãn. Một số phương pháp tiếp cận phổ biến của phân lớp câu chủ quan là: dùng phương pháp Naïve Bayes (đã trình bày ở mục trước), phương pháp phân lớp sử dụng mẫu [16].

Trong việc phân lớp câu chủ quan và khách quan cho Tiếng Việt thể kể đến như mô hình tự động học trong phân loại chủ quan Tiếng Việt [17]. Việc phân lớp chủ quan trong nghiên cứu này được thực hiện qua phương pháp sử dụng mẫu nhưng điểm khác biệt là các thông tin POS được chọn làm đặc trưng cho các mẫu huấn luyện.



Hình 2.7 Mô hình phân lớp chủ quan cho Tiếng Việt.

Quy trình thực hiện trải qua 2 bước sau:

- Bước 1: Trích xuất tất cả mẫu dữ liệu huấn luyện đã được gán nhãn.
- Bước 2: Đánh giá các mẫu để chọn bộ mẫu tốt nhất.

Trong việc chọn ra bộ mẫu tốt nhất nhóm tác giả đã thực hiện 2 giai đoạn, giai đoạn 1 chọn ra bộ mẫu ở mức chấp nhận được dựa vào tần suất xuất hiện trong các dữ liệu chủ qua so với khách quan. Ở giai đoạn 2 để chọn ra được bộ mẫu tốt nhất từ bộ mẫu chấp nhận được dựa vào các đặc trưng POS trong dữ liệu theo quy tắc được đặt ra [17].

2.4 Biểu diễn văn bản

Biểu diễn văn bản là một bước quan trọng trong khai thác dữ liệu văn bản, truy vấn thông tin và xử lý ngôn ngữ tự nhiên. Các mô hình biểu diễn đóng vai trò trung gian giữa ngôn ngữ tự nhiên dạng văn bản và các chương trình xử lý.

Văn bản ở dạng thô (chuỗi) sau khi được chuyển sang mô hình sẽ trở thành những cấu trúc dữ liệu trực quan, đơn giản hơn, thuận lợi cho việc hiểu và tính toán trên văn bản. Vì vậy, các mô hình biểu diễn văn bản không ngừng cải thiện và phát triển. Tùy thuộc vào từng bài toán, từng thuật toán khác nhau mà chúng ta có mô hình biểu diễn phù hợp.

Các mô hình biểu diễn văn bản truyền thống như mô hình túi từ (bag-of-word), mô hình không gian vector là các mô hình thường được sử dụng nhất. Tuy nhiên, các mô hình này có nhược điểm là không nắm bắt được các thông tin cấu trúc quan trong của văn bản như trật tự các từ, vị trí của từ trong văn bản. Mô hình đồ thị biểu diễn văn bản, word2vec, sentence2vec là phương pháp mới đang được quan tâm và sử dụng trong các lĩnh vực khai phá dữ liệu văn bản hiện tại.

2.4.1 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 biểu diễn [18].

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:

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)
Công	VB1(6)
Hoàn	VB2(3)
Nhiệm	VB2(5)
Vụ	VB2(6)

Bảng 2.1 Biểu diễn văn bản trong mô hình logic

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 vần chữ cái. Phép tìm kiếm trên có độ phức tạp cấp $(\mathrm{nlog}_2\, n)$, 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)

▪ **Hạn chế:** 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.

2.4.2 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 [18].

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.

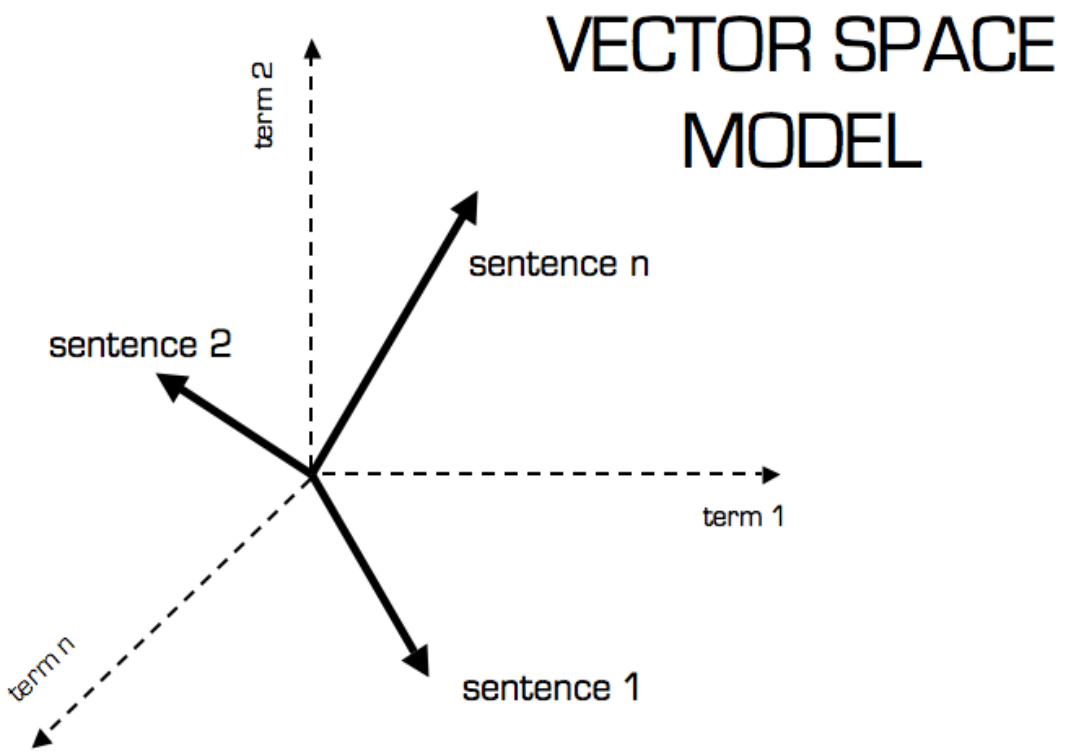
Ưu điểm: Tìm kiếm theo phương pháp này khá hiệu quả và đơn giản, do tìm kiếm nhanh và chính xác. Đối với những ngôn ngữ đơn giản về mặt ngữ pháp thì việc phân tích trên có thể đạt được mức độ chính xác cao và chấp nhận được.

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.

2.4.3 Mô hình không gian vector

Mô hình vector là một trong những mô hình đơn giản và thường được sử dụng trong phần lớn các bài toán xử lý dữ liệu văn bản. Nói một cách ngắn gọn, mô hình không gian vector (Vector space model) là một mô hình đại số (algebraic model) thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện của nó trong một tài liệu.

Mô hình này biểu diễn văn bản như những điểm trong không gian Euclid nchiều, mỗi chiều tương ứng với một từ trong tập hợp các từ. Phần tử thứ i, là di của vector văn bản cho biết số lần mà từ thứ i xuất hiện trong văn bản. Sự tương đồng của hai văn bản được định nghĩa là khoảng cách giữa các điểm, hoặc là góc giữa những vector trong không gian.



Hình 2.8 Mô hình không gian vector.

Giả sử ta có một văn bản và nó được biểu diễn bởi vector $\vec{v}(v_1, v_2 \dots, v_n)$. Trong đó n là số đặc trưng hay số chiều của vector (thường là số từ khóa), v_i là trọng số của đặc trưng thứ i (với $1 \leq i \leq n$).

Ví dụ: xét 2 văn bản với trọng số đặc trưng là số lần xuất hiện của từ khóa thứ i trong văn bản, vector biểu diễn tương ứng như sau:

VB1: Máy vi tính

VB2: Siêu máy tính

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

Từ	**Vector VB1**	**Vector VB2**
Máy	1	1
vi	1	0
tính	1	1

Trọng số của đặc trưng có thể tính dựa trên tần số xuất hiện của từ khóa trong văn bản. Ma trận biểu diễn trọng số (ma trận tần suất) $w = \{w_{ij}\}$ được xác định dựa trên tần số xuất hiện của từ khóa t_i trong văn bản d_j. Một số phương pháp xác định w_{ij}:

- Phương pháp Boolean weighting: giá trị là 1 nếu số lần xuất hiện của từ khóa lớn hơn một ngưỡng nào đó, ngược lại 0.
- Phương pháp dựa trên tần số từ khóa (Term Frequency Weighting).
- Phương pháp dựa trên nghịch đảo tần số văn bản (Inverse Document Frequency).
- TF*IDF weighting.

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.

2.4.3.1 Phương pháp 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 t_i sẽ cho ra giá trị đúng nếu và chỉ nếu từ khóa t_i xuất hiện trong văn bản đó.

Mô hình Boolean được xác định như sau:

$$W_{ij} = \begin{cases} 1 & nếu\ \ t_i\ có\ trong\ d_j \\ 0 & ngược\ lại \end{cases}$$

Giả sử có một cơ sở dữ liệu gồm $\boldsymbol{m}$ văn bản $D = \{d_1, d_2 \dots, d_m\}$. Mỗi văn bản được biểu diễn dưới dạng một vector gồm $\boldsymbol{n}$ từ khóa $T = \{t_1, t_2 \dots, t_n\}$. Gọi $W = \{W_{ij}\}$ là ma trận trọng số, trong đó W_{ij} là giá trị trọng số của từ khóa t_i trong văn bản d_j.

2.4.3.2 *Phương pháp dựa trên tần số từ khóa (Term Frequency)*

TF: Tần suất thời gian, đo tần suất một thuật ngữ xuất hiện thường xuyên trong một văn bản. Vì mỗi văn bản đều khác nhau về chiều dài, có thể một thuật ngữ sẽ xuất hiện nhiều hơn trong các văn bản dài hơn và nó sẽ xuất hiện ít hơn trong các văn bản ngắn hơn. Do đó, tần suất cụm từ thường được chia cho độ dài văn bản (còn gọi là tổng số thuật ngữ trong văn bản) như một cách chuẩn hóa: TF (t) = (Số lần t xuất hiện trong văn bản) / (Tổng số các thuật ngữ trong văn bản).

Giá trị trọng số từ khóa W_{ij} được tính dựa trên tần số xuất hiện của từ khóa trong văn bản. Giả sử f_{ij} là số lần xuất hiện của từ khóa t_i trong văn bản d_j, khi đó W_{ij} được tính bởi một trong ba công thức:

$$W_{ij} = f_{ij}$$

$$W_{ij} = 1 + \log f_{ij}$$

$$W_{ij} = \sqrt{f_{ij}}$$

Nếu số lần xuất hiện từ khóa ti trong văn bản d_j càng lớn thì có nghĩa là văn bản dj càng phụ thuộc vào từ khóa t_i, hay nói cách khác từ khóa t_i mang nhiều thông tin trong văn bản d_j. Ví dụ nếu trong văn bản xuất hiện nhiều từ khóa giảng viên học sinh, điều đó có nghĩa là văn bản chủ yếu liên quan đến lĩnh vực giáo dục.

2.4.3.3 *Phương pháp dựa trên nghịch đảo tần số văn bản*

IDF: Tần số nghịch của 1 từ trong tập văn bản, đo lường mức độ quan trọng của một thuật ngữ. Trong khi tính toán TF tất cả các từ được coi là quan trọng không kém. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn như:

- Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy, …
- Giới từ: ở, trong, trên, …
- Từ chỉ định: ấy, đó, nhỉ, …

Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF :

IDF (t) = log_e (Tổng số văn bản / Số văn bản có thời hạn t trong đó).

Trong phương pháp này W_{ij} được tính theo công thức sau:

$$W_{ij} = \begin{cases} \log(\frac{N}{df_i}) & nếu\ tf_i \geq 1 \\ 0 & nếu\ tf_i = 0 \end{cases}$$

Trong đó N là số lượng văn bản và df_i là số lượng văn bản mà từ khóa t_i xuất hiện. Trong công thức này, trọng số W_{ij} được tính dựa trên độ quan trọng của từ khóa t_i trong văn bản d_j. Nếu t_i xuất hiện trong càng ít văn bản, thì khi nó xuất hiện trong d_j nào thì trọng số của nó đối với d_j càng lớn (do tính nghịch đảo của hàm log), tức là hàm lượng thông tin trong nó càng lớn. Nói cách khác t_i là điểm quan trọng để phân biệt d_j với các văn bản khác.

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

2.4.3.4 Phương pháp TF – IDF

TF-IDF (Term Frequency-Inverse Document Frequency) là phương pháp kết hợp của hai phương pháp TF và IDF. Trọng số này sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Trọng số W_{ij} được tính bằng tần số xuất hiện của từ khóa $t_i$trong văn bản d_j và độ quan trọng của từ khóa d_j trong tập văn bản.

Công thức tính W_{ij} là:

$$W_{ij} = \begin{cases} (1 + \log f_{ij}) \log(\frac{N}{df_i}) & nếu\ f_{ij} \geq 1 \\ 0 & nếu\ f_{ij} = 0 \end{cases}$$

Trong đó:

- f_{ij} (term frequency): số lần xuất hiện của từ t_i trong văn bản thứ d_j, f_{ij} càng cao thì từ đó càng miêu tả tốt nội dung văn bản.
- df_i (document frequency): số văn bản có chứa từ t_i.

Nhận xét về mô hình không gian vector:

- Ưu điểm: mô hình vector là mô hình biểu diễn văn bản được sử dụng khá phổ biến trong các hệ xử lý văn bản. Mối quan hệ giữa các văn bản được tính toán dựa trên các vector biểu diễn nên dễ dàng thực hiện.
- Nhược điểm: vì mỗi văn bản được biểu diễn thành một vector n chiều, với số chiều thường là số từ khác nhau trong tập văn bản, do đó không gian biểu diễn có số chiều tương đối lớn, việc lưu trữ và tính toán trên vector tốn kém và phức tạp. Ngoài ra, hệ thống không linh hoạt khi lưu trữ các từ khóa. Chỉ cần một thay đổi nhỏ trong bảng từ vựng sẽ dẫn đến hoặc là vector hóa lại toàn bộ các tài liệu, hoặc là bỏ qua các từ có nghĩa bổ sung trong các tài liệu được mã hóa trước đó.

2.4.4 Mô hình đồ thị

Hiện nay, trên thế giới có một số công trình xử lý văn bản sử dụng mô hình đồ thị. Các mô hình đồ thị tương đối đa dạng và mỗi mô hình mang nét đặc trưng riêng. Mỗi đồ thị là một văn bản hoặc biểu diễn cho tập văn bản. Đỉnh của đồ thị có thể là câu, hoặc từ, hoặc kết hợp câu và từ. Cạnh nối giữa các đỉnh là vô hướng hoặc có hướng, thể hiện mối quan hệ trong đồ thị. Nhãn đỉnh thường là tần số xuất hiện của đỉnh. Còn nhãn cạnh là tên mối liên kết khái niệm giữa 2 đỉnh, hay tần số xuất hiện chung của 2 đỉnh trong một phạm vi nào đó, hay tên vùng mà đỉnh xuất hiện.

Mô hình đồ thị biểu diễn văn bản cụ thể là mô hình đồ thị khái niệm (Conceptual Graphs_ CGs) được John F. Sowa trình bày lần đầu tiên vào năm 1976 [19]. Hiện nay, mô hình đồ thị không ngừng phát triển dựa trên ý tưởng của mô hình CGs và được ứng dụng rộng rãi vào các bài toán liên quan đến xử lý văn bản.

Ví dụ, trong bài toán rút trích thông tin, đỉnh là từ hay từ kết hợp câu, cạnh thể hiện tần số xuất hiện. Trong bài toán phân lớp văn bản, đỉnh là từ, cạnh thể hiện trật tự xuất hiện của từ hay vị trí xuất hiện của từ trong văn bản. Trong bài toán tóm tắt văn bản, đỉnh là câu, cạnh thể hiện sự tương đồng giữa các câu.

Các dạng mô hình đồ thị:

- Mô hình đồ thị sử dụng đỉnh là từ trong văn bản: gồm mô hình đồ thị sử dụng mạng ngữ nghĩa và mô hình đồ thị không sử dụng mạng ngữ nghĩa.
- Mô hình đồ thị sử dụng đỉnh là câu.
- Mô hình đồ thị đỉnh là câu và từ.

2.4.5 Mô hình túi từ (Bag of word)

Mô hình túi từ ngữ (Bag-of-Word - BOW) là một biểu diễn đơn giản hóa của văn bản thường được sử dụng trong xử lý ngôn ngữ tự nhiên và tìm kiếm thông tin [20]. Ý tưởng của BoW là phân tích và phân nhóm dựa theo túi từ ngữ. Mô hình túi từ ngữ học được một bộ từ vựng từ tất cả các văn bản, rồi mô hình các văn bản bằng cách đếm số lần xuất hiện của mỗi từ trong văn bản đó.

Ví dụ, với hai câu sau :

Câu 1: “Nam là học sinh giỏi toán” .

Câu 2: “Nam hướng dẫn Hoa học toán để cùng nhau học giỏi toán” .

Sau khi tiền xử lý hai câu trên trở thành :

Câu 1: {Nam, học_sinh, giỏi, toán} .

Câu 2: {Nam, hướng_dẫn, Hoa, học, toán, cùng_nhau, học, giỏi, toán} .

Hai câu trên có tập từ là:

{Nam, học_sinh, giỏi, toán, hướng_dẫn, Hoa, học, cùng_nhau}

Để có thể xây dựng được bag of words của 2 câu này, chúng ta đếm số lần xuất hiện của mỗi từ trong mỗi câu.

Trong câu 1 “Nam” , “học_sinh”, “giỏi”, “toán” xuất hiện một lần. Như thế đặc tính vector cho câu 1 sẽ là:

{ 1, 1, 1, 1, 0, 0, 0, 0 }

Trong câu 2 “học” với “toán” xuất hiện hai lần và “Nam”, “hướng_dẫn ”, “Hoa”, “cùng_nhau”, “giỏi” xuất hiện một lần. Như thế đặc tính vector cho câu 2 sẽ là :

{ 1, 0, 1, 2, 1, 1, 2, 1}

Trong mô hình túi từ ngữ một văn bản được biểu diễn như một tập hợp (gọi là "túi") các từ xuất hiện trong văn bản, không quan tâm đến ngữ pháp và thứ tự xuất hiện của các từ mà chỉ lưu lại tần suất xuất hiện của mỗi từ trong văn bản.

Mô hình túi từ ngữ thường được sử dụng trong các phương pháp phân loại văn bản khi mà tần suất xuất hiện của từ được sử dụng như là một đặc trưng để huấn luyện một bộ phân lớp.

2.4.6 Mô hình Word2vec

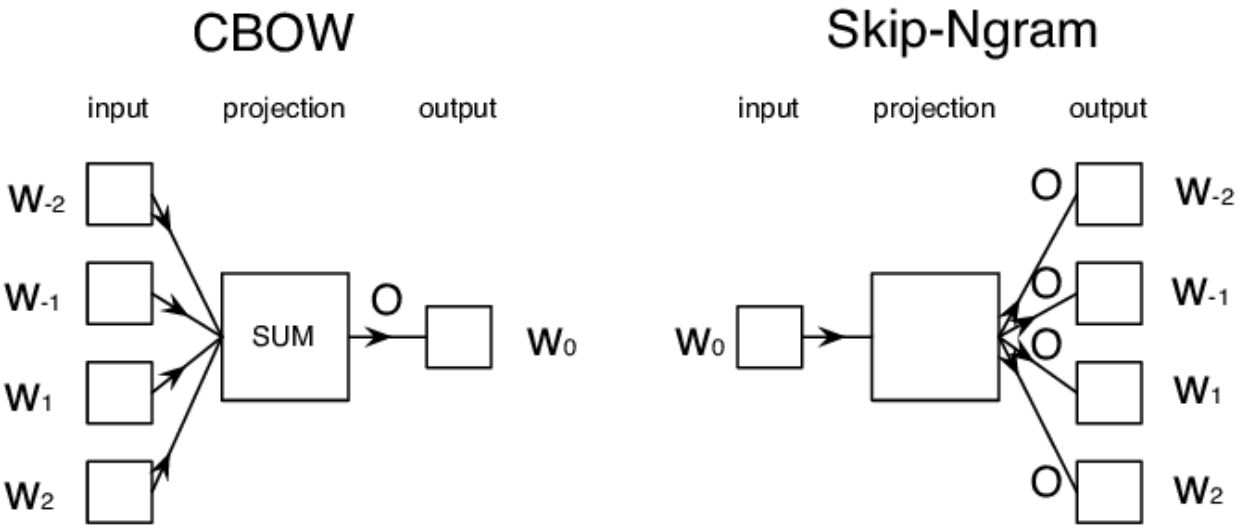
Word2vec được giới thiệu vào năm 2013 bởi Tomas Mikolov. Nó là một mạng neural hai lớp xử lý văn bản. Đầu vào của nó là một phần văn bản và đầu ra của nó là một tập các vectơ đặc trưng cho các từ trong kho văn bản đó.

Trong word2vec, một biểu diễn phân tán của một từ được sử dụng. Tạo ra một vector với kích thước rất nhiều chiều. Mỗi từ được biểu diễn bởi tập các trọng số của từng phần tử trong nó. Vì vậy, thay vì sự kết nối một một giữa một phần tử trong vector với một từ, biểu diễn từ sẽ được dàn trải trên tất cả các thành phần trong vector, và mỗi phần tử trong vector góp phần định nghĩa cho nhiều từ khác nhau.

Mỗi vector như vậy cũng đại diện cho một cách tóm lược của ý nghĩa của một từ. Chỉ đơn giản bằng cách kiểm tra một ngữ liệu lớn, nó có thể tự động học word vectors và nắm bắt các mối quan hệ giữa các từ.

Word2vec bao gồm 2 mô hình [21]:

- Mô hình túi từ liên lục (CBOW): dự đoán 1 từ khi đã có các từ lân cận. CBOW có điểm thuận lợi là training mô hình nhanh hơn so với mô hình skip-gram, thường cho kết quả tốt hơn với frequence words (hay các từ thường xuất hiện trong văn cảnh).
- Mô hình Skip-gram: là một mô hình đối lập hoàn toàn với mô hình CBOW. Mô hình này giúp dự đoán các từ lân cận khi đã có 1 từ. Skip-gram huấn luyện chậm hơn.

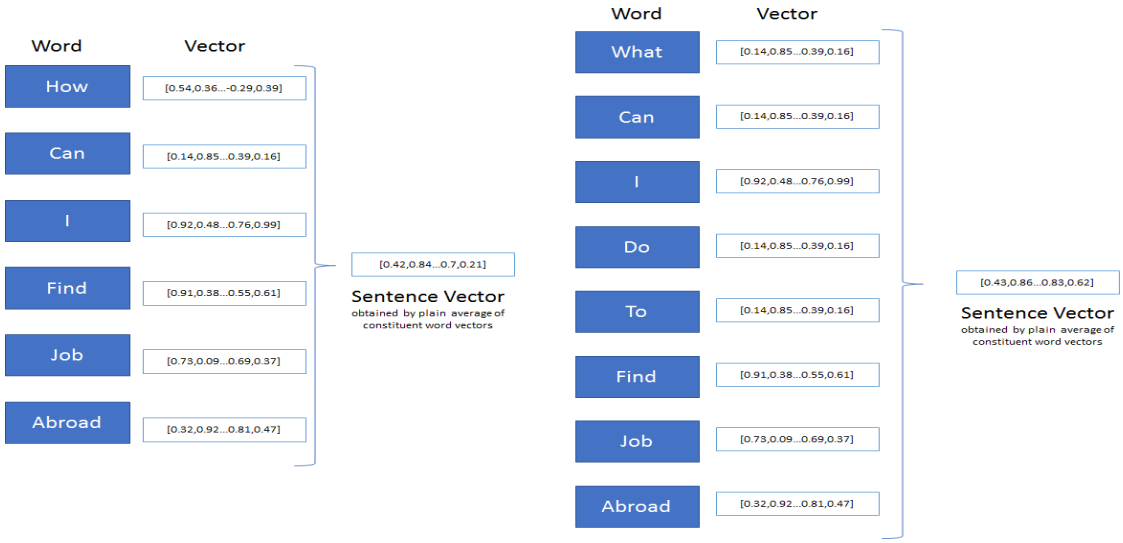


Hình 2.9 Mô hình CBOW và Skip-gram trong Word2vec.

Mục đích và tính hữu ích của word2vec là nhóm các vectơ của các từ tương tự lại với nhau trong vectorspace. Nghĩa là, nó phát hiện các điểm tương đồng về mặt toán học

2.4.7 Mô hình Sentence2vec

Mô hình Sentence2vec là phương pháp mô hình hóa câu văn lên không gian vector. Như ở phần trước đã tìm hiểu Word2vec là phương pháp biểu diễn vector cho từ, mỗi từ cũng được biểu diễn thành vector trọng số nhiều chiều, vậy chúng ta có thể hiểu đơn giản để xây dựng mô hình sentence2vec cho câu chúng ta có thể lấy tổng hoặc lấy tổng trung bình vector của các từ cấu thành lên câu để làm vector biểu diễn cho câu.



Hình 2.10 Mô hình biểu diễn sentence2vec.

2.5 Các phương pháp tính độ tương đồng văn bản

Độ tương đồng là một đại lượng dùng để so sánh hai hay nhiều đối tượng với nhau, phản ánh cường độ của mối quan hệ giữa các đối tượng với nhau. Ví dụ: xét 2 câu "Nam là sinh viên lớp công nghệ thông tin " và "Hoa là sinh viên lớp công nghệ thông tin", ta có thể nhận thấy hai câu trên có sự tương đồng cao.

Phát biểu bài toán tính độ tương đồng như sau: Xét 2 văn bản d_i và d_j. Mục tiêu là tìm ra một giá trị S(d_i, d_j), S € (0,1), thể hiện độ tương đồng giữa 2 văn bản d_i và d_j. Giá trị càng cao thì sự giống nhau về nghĩa của hai văn bản càng nhiều. Ví dụ trong mô hình không gian vector người ta sử dụng độ đo Cosine để tính độ tương đồng giữa hai văn bản, mỗi văn bản được biểu diễn bởi một vector. Độ tương tự ngữ nghĩa là khái niệm thể hiện tỷ lệ dựa trên sự giống nhau về nội dung ý nghĩa của tập các tài liệu hoặc các thuật ngữ trong một danh sách các thuật ngữ. Độ tương đồng ngữ nghĩa phản ánh mối quan hệ ngữ nghĩa giữa các câu, các tài liệu văn bản.

Độ tương tự giữa các câu đóng một vai trò ngày càng quan trọng trong nghiên cứu về khai thác dữ liệu và xử lý ngôn ngữ tự nhiên. Nó cũng được sử dụng như là một tiêu chuẩn của trích chọn thông tin để tìm ra những tri thức ẩn trong cơ sở dữ liệu hay trên các kho dữ liệu trực tuyến [18].

Một số phương pháp tính độ tương đồng câu hiện nay:

- Tính độ tương đồng dựa trên tập từ chung.
- Tính độ tương đồng dựa trên vector biểu diễn.
- Tính độ tương đồng dựa trên ngữ nghĩa.
- Tính độ tương đồng dựa trên thứ tự từ.

Ở nghiên cứu này tôi sẽ tập trung nghiên cứu một số phương pháp tính độ tương đồng dựa trên vector biểu diễn như: dựa vào khoảng cách Cosine, dựa vào khoảng cách Manhattan, dựa vào khoảng cách Euclide.

2.5.1 Độ tương đồng Cosine

Trong phương pháp này, các văn bản được biểu diễn theo mô hình không gian vector, mỗi thành phần của vector chỉ đến một từ tương ứng trong danh sách mục từ đã thu được từ quá trình tiền xử lý văn bản đầu.

Không gian vector hay số chiều của vector có kích thước bằng số mục từ trong danh sách mục từ. Giá trị mỗi phần tử của vector là độ quan trọng của mục từ trong câu.

Độ quan trọng của từ được tính theo một trong các phương pháp đã trình bày ở phần trên

Giả sử vector biểu diễn cho hai văn bản lần lượt có dạng:

$D_i = \{w_1^i, w_2^i, \dots w_t^i\}$ với w_t^i là trọng số của từ thứ t trong không gian vector i.

$D_i = \{w_1^j, w_2^j, \dots w_t^j\}$ với w_t^i là trọng số của từ thứ t trong không gian vector j.

Độ đo tương đồng được tính là Cosine của góc giữa hai vector biểu diễn cho hai văn bản Di và Dj. Độ tương tự của chúng được tính theo công thức :

$$sim(D_{ij}) = \frac{\sum_{k=1}^{t} w_k^i w_k^j}{\sum_{k=1}^{t}(w_k^i)^2 * \sum_{k=1}^{t}(w_k^j)^2}$$

Nhận xét: vector biểu diễn cho các câu chưa quan tâm đến mối quan hệ ngữ nghĩa giữa các từ mục do đó các từ đồng nghĩa sẽ không được phát hiện, kết quả so sánh độ tương tự giữa hai văn bản chưa có sự chuẩn xác cao.

2.5.2 Độ tương đồng Manhattan

Khoảng cách Manhattan là phương pháp tính độ tương đồng giữa các vector đặc trưng biểu diễn cho hai văn bản .

Cho hai vector $\overrightarrow{v_a}$ và $\overrightarrow{v_b}$, khoảng cách Manhattan được định nghĩa như sau:

$$man\_dist_{(\overrightarrow{v_a},\overrightarrow{v_b})} = \sum_{i=1}^{n} |w_{ai} - w_{ai}|$$

Mức độ tương đồng giữa hai vector được xác định bằng công thức:

$$man\_sim = 1 - \frac{man_{dist(\overrightarrow{v_a},\overrightarrow{v_b})}}{n} = 1 - \frac{1}{n}\sum_{i=1}^{n} |w_{ai} - w_{ai}|$$

2.5.3 Độ tương đồng Euclide

Khoảng cách Euclide cũng là một phương pháp khá phổ biến để xác định mức độ tương đồng giữa các vector đặc trưng của hai văn bản. Cho hai vector a và b, khoảng cách Euclide được định nghĩa như sau:

$$e\_dist_{(\overrightarrow{v_a},\overrightarrow{v_b})} = \sqrt{\sum_{i=1}^{n}(w_{ai} - w_{ai})^2}$$

Mức độ tương đồng giữa hai vector được xác định bằng công thức:

$$e\_sim = 1 - \frac{e\_dist_{(\overrightarrow{v_a},\overrightarrow{v_b})}}{n} = 1 - \frac{1}{n}\sqrt{\sum_{i=1}^{n}(w_{ai} - w_{ai})^2}$$

2.6 Các phương pháp tiền xử lý văn bản

Văn bản trước khi đưa vào mô hình xử lý cần được tiền xử lý. Quá trình này sẽ giúp nâng cao hiệu quả của mô hình và giảm độ phức tạp của thuật toán được cài đặt vì nó có nhiệm vụ làm giảm số từ có trong biểu diễn văn bản. Các bước xử lý văn bản gồm: tách từ, loại bỏ từ có tần số thấp và xác định từ đồng nghĩa [18].

2.6.1 Tách từ

Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ là 1 trong những bài toán cơ bản và quan trọng bậc nhất. Ví dụ: từ "đất nước" được tạo ra từ 2 âm tiết "đất" và "nước", cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành một bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, so sánh văn bản, tóm tắt văn bản, máy dịch tự động.

Tách từ chính xác hay không là công việc rất quan trọng, nếu không chính xác rất có thể dẫn đến việc ý nghĩa của câu sai, ảnh hưởng đến tính chính xác của chương trình. Bước này có nhiệm vụ xác định các từ có trong văn bản, kết quả của nó là một tập các từ riêng biệt. Các trường hợp đặc biệt như số, dấu ngoặc, dấu chấm câu thường bị loại ra trong khi phân tích vì một mình nó không mang lại ý nghĩa nào cho tài liệu (ngoại trừ một vài trường hợp đặc biệt, ví dụ trong thu thập thông tin về lĩnh vực lịch sử). Tuy nhiên trong một vài trường hợp, chẳng hạn đối với những từ ghép nối (state-of-the-art) không được phép bỏ dấu "-", vì sẽ làm thay đổi nghĩa của từ [23].

Đã có nhiều công trình nghiên cứu xây dựng mô hình tách từ tiếng Việt và đạt được những kết quả chính xác cao như mô hình tách từ bằng WFST (Weighted Finit State Transduce) và mạng Neural đã được sử dụng trong công trình của tác giả Đình Điền (2001). Công cụ tách từ JvnTextPro do nhóm tác giả Nguyễn Cẩm Tú, Khoa Công nghệ - Trường Đại học Quốc gia Hà Nội. Bộ công cụ tách từ vnTokenizer của tác giả Lê Hồng Phương. Nhiều hướng tiếp cập trong bài toán tách từ được đưa ra, trong nhiên cứu của Đỗ Thị Thanh Nga, "Tính toán độ tương tự ngữ nghĩa văn bản dựa vào độ tương tự giữa từ với từ" tác giả đã chỉ ra sơ đồ bài toán tách từ gồm hai hướng đó là dựa trên từ và dựa trên kí tự.

Các hướng tiếp cận dựa trên "từ": hướng tiếp cận dựa trên từ với mục tiêu tách được các từ hoàn chỉnh trong câu.

- **Hướng tiếp cận dựa trên thống kê:** Dựa trên các thông tin thống kê như tần số xuất hiện của từ trong tập huấn luyện ban đầu. Hướng tiếp cận này đặc biệt dựa trên tập ngữ liệu huấn luyện. Nhờ vậy, hướng tiếp cận này tỏ ra linh hoạt và hữu dụng trong nhiều lĩnh vực khác nhau.
- **Hướng tiếp cận dựa trên từ điển:** Ý tưởng của hướng tiếp cận này là những cụm từ được tách ra từ văn bản phải được so khớp với các từ trong từ điển. Do đó trong hướng tiếp cận này đòi hỏi từ điển riêng cho từng lĩnh vực quan tâm.
- **Hướng tiếp cận theo Hybrid:** Với mục đích kết hợp các hướng tiếp cận khác nhau để thừa hưởng được các ưu điểm của nhiều kỹ thuật và các hướng tiếp cận khác nhau nhằm nâng cao kết quả. Hướng tiếp cận này thường kết hợp giữa hướng dựa trên thống kê và dựa trên từ điển nhằm tận dụng các mặt mạnh của các phương pháp này. Tuy nhiên hướng tiếp cận Hybrid lại mất nhiều thời gian xử lý, không gian đĩa và đòi hỏi nhiều chi phí.

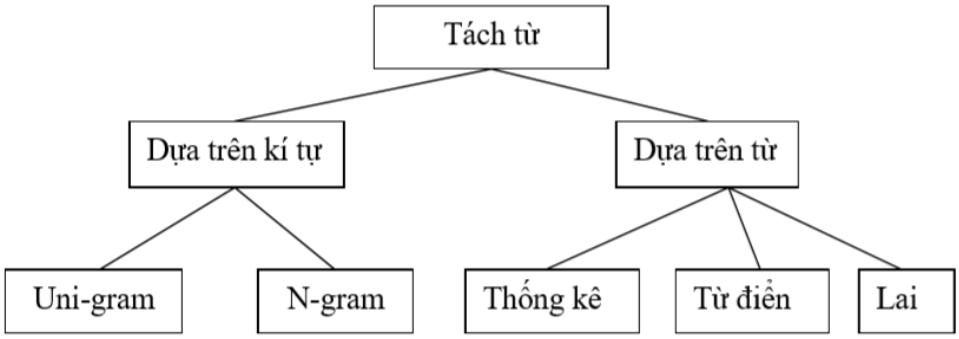
Các hướng tiếp cận dựa trên ký tự:

Các hướng tiếp cận dựa trên ký tự (dựa trên "tiếng" trong tiếng Việt) có thể chia làm 2 nhóm nhỏ: uni-gram và n-gram.

Trong tiếng việt, hình vị nhỏ nhất là "tiếng" được hình thành bởi nhiều ký tự trong bảng chữ cái. Hướng tiếp cận này đơn thuần rút trích ra một số lượng nhất định các tiếng

trong văn bản như rút trích từ 1 ký tự (uni-gram) hay nhiều ký tự (ngram). Theo tác giả nghiên cứu thì hướng tiếp cận dựa trên nhiều ký tự có nhiều ưu điểm nổi bật hơn. Nó đơn giản, dễ ứng dụng, ngoài ra còn có thuận lợi là ít tốn chi phí cho thao tác tạo chỉ mục và xử lý nhiều câu truy vấn. Qua nhiều công trình nghiên cứu của các tác giả đã được công bố, hướng tiếp cận tách từ dựa trên nhiều ký tự, cụ thể là cách tách từ hai ký tự được cho là sự lựa chọn thích hợp.

Một số phương pháp tách từ tiếng Việt hiện nay: Phương pháp Maximum Matching: Forward/Backward, Phương pháp Transformation-based Learning (TBL), Mô hình tách từ bằng WFST và mạng Neural



Hình 2.11 Các phương pháp tiếp cận trong tách từ.

2.6.2 Loại bỏ từ dừng

Từ dừng là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng Việt từ dừng là những từ như: “như vậy”, “sau đó”, “một số”, “chỉ”, “của” .v.v

Có rất nhiều cách để loại bỏ từ dừng nhưng có 2 cách chính là: dùng từ điển và dựa theo tần suất xuất hiện của từ.

Với phương pháp dùng từ điển cách này đơn giản nhất, chúng ta tiến hành lọc văn bản, loại bỏ những từ xuất hiện trong từ điển StopWords. Đối với phương pháp dựa theo tần suất xuất hiện của từ chúng ta tiến hành đếm số lần xuất hiện của từng từ trong data sau đó sẽ loại bỏ những từ xuất hiện nhiều lần (cũng có thể là ít lần). Khoa học đã chứng minh những từ xuất hiện nhiều nhất thường là những từ không mang nhiều ý nghĩa.

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

3.1Giới thiệu

Bài toán phân loại văn bản là một bài toán rất phổ biến trong xử lý ngôn ngữ tự nhiên hiện nay, ví dụ bài toán phân loại cảm xúc hay thái độ của người dùng qua bình luận (comment) trên các trang phim, đánh giá về sản phẩm … Hay như trong ứng dụng chatbot, bài toán phân loại văn bản được sử dụng để phát hiện mục đích của người dùng.

Dựa vào việc phân loại được tự động các bình luận chúng ta có thể đánh giá được chất lượng của một sản phẩm, dịch vụ, xu hướng lên khách hàng, cộng đồng là tích cực hay tiêu cực để có những chiến lược kinh doanh phù hợp. Các công cụ như thế kết hợp với các công cụ thu thập dữ liệu tự động từ nhiều nguồn khác nhau (mạng xã hội, báo điện tử, diễn đàn…) sẽ tạo lên bộ công cụ điều tra thăm dò cực kỳ giá trị.

Có thể hiểu phân loại cảm xúc là quá trình dự đoán và gán văn bản vào một hoặc nhiều cảm xúc trước đó, ở mức độ đơn giản sẽ là ở 2 cảm xúc tích cực (positive) và tiêu cực (negative). Phân loại cảm xúc tự động là một lĩnh vực nghiên cứu được quan tâm trong nhiều năm qua do khả năng ứng dụng rộng rãi và hiệu quả sử dụng. Những phương pháp phổ biến được sử dụng để thực hiện việc phân loại như là: Naïve Bayes, k-láng giềng gần nhất (k-NN), mạng nơron, máy vectơ hỗ trợ (SVM), các phương pháp này đều sử dụng mô hình không gian vectơ khi biểu diễn văn bản.

Mô hình không gian vectơ là phương pháp biểu diễn văn bản phổ biến. Trong đó, mỗi từ trong văn bản có thể trở thành đặc trưng (hay chiều của vectơ biểu diễn văn bản). Đây là mô hình biểu diễn văn bản rất phổ biến và được sử dụng rộng rãi.

Bên cạnh đó trong quá trình công tác tại Trường Đại học Công Nghệ Tp.HCM tôi nhận thấy việc thu thập ý kiến khảo sát sinh viên về chất lượng giảng dạy của giảng viên ở mỗi học kỳ là một vấn đề thực tiễn và có thể áp dụng mô hình phân loại cảm xúc tự động cho những ý kiến đánh giá của sinh viên.

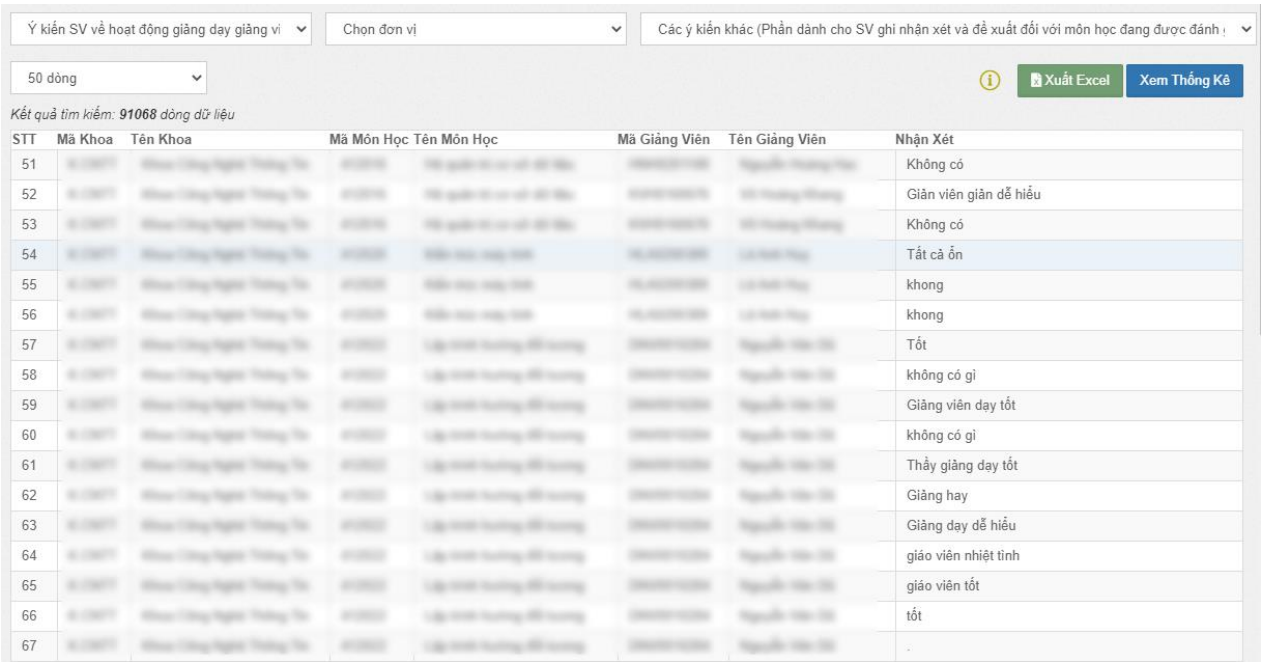
Trên những cơ sở đó trong luận văn này tôi sẽ nghiên cứu và trình bày một giải pháp để xây dựng mô hình phân loại ý kiến đánh giá tự động trong môi trường giáo dục. Luận văn cũng sẽ nghiên cứu và so sánh độ hiệu quả giữa các phương pháp phân loại khác nhau

trên cùng một tập dữ liệu để có thể làm nguồn tài liệu tham khảo cho những nghiên cứu có tập dữ liệu tương tự.

3.2Thu thập dữ liệu

Dữ liệu tôi sử dụng trong luận văn này được lấy từ nguồn dữ liệu khảo sát ý kiến sinh viên về chất lượng giảng dạy của giảng viên trong học kì I năm học 2016-2017 của Trường Đại học Công Nghệ Tp.HCM.

Dữ liệu này được thu thập thông qua cổng thông tin trực tuyến của nhà trường và được trích xuất toàn bộ và chưa qua bất kỳ bộ xử lý nào.



STT	Mã Khoa	Tên Khoa	Mã Môn Học	Tên Môn Học	Mã Giảng Viên	Tên Giảng Viên	Nhận Xét
51							Không có
52							Giản viên giản dễ hiểu
53							Không có
54							Tất cả ổn
55							khong
56							khong
57							Tốt
58							không có gì
59							Giảng viên dạy tốt
60							không có gì
61							Thầy giảng dạy tốt
62							Giảng hay
63							Giảng dạy dễ hiểu
64							giáo viên nhiệt tình
65							giáo viên tốt
66							tốt
67							.

Hình 3.1 Hình chụp dữ liệu khảo sát ý kiến sinh viên

Do nguồn dữ liệu này chưa được xử lý nên tập dữ liệu tồn tại những hạn chế nhất định như: chứa nhiều câu quá ngắn, câu sai chính tả, câu sử dụng ngôn ngữ khác Tiếng Việt. Vì vậy trước khi có thể đưa vào áp dụng thực tế trong bài toán của luận văn này đòi hỏi dữ liệu phải được chọn lọc ở mức cơ bản. Trong tập dữ liệu được sử dụng này đã được tôi loại bỏ những dữ liệu:

- Dữ liệu quá ngắn (những ý kiến dưới 5 từ).
- Dữ liệu sai chính tả trên 30% câu.
- Dữ liệu sử dụng các ngôn ngữ khác ngoài Tiếng Anh

- Dữ liệu trung tính (neutral)

Sau những bước xử lý chọn lọc dữ liệu tôi đã chọn ra được 1100 dữ liệu ý kiến đánh giá để sử dụng trong luận văn này. Toàn bộ dữ liệu này sau đó sẽ được gán nhãn thủ công vào 2 nhãn tích cực (positive) và tiêu cực (negative).

3.3Quy trình thực hiện

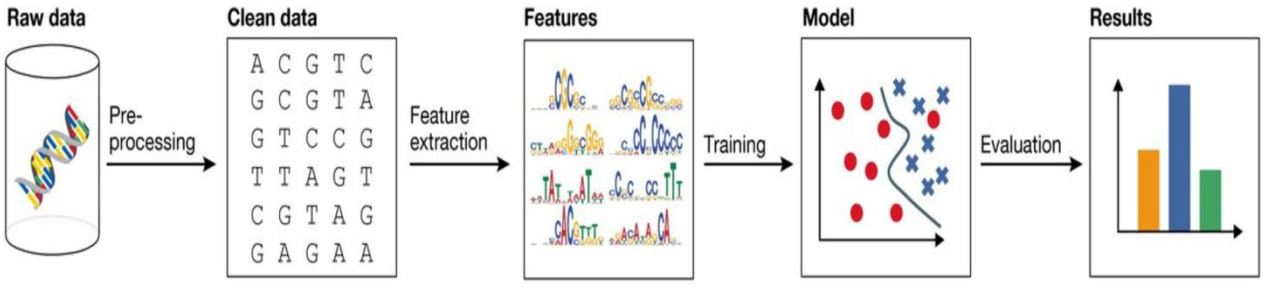
3.3.1 Tiền xử lý văn bản

Tiền xử lý văn bản được xem là một bước không thể thiếu trong việc xây dựng một bộ phân lớp nhằm cải thiện độ chính xác phân lớp. Vì văn bản vốn được thu thập liệt kê mà không có cấu trúc nếu giữ nguyên sẽ rất khó khăn trong xử lý. Đặc biệt là các loại văn bản được thu thập từ các nguồn website sẽ lẫn chứa các HTML code, code đây gọi là nhiễu dữ liệu và xử lý làm sạch dữ liệu.

Về cơ bản tiền xử lý văn bản sẽ bao gồm các bước:

- Làm sạch văn bản
- Tách từ:
- Chuẩn hóa từ
- Loại bỏ stopwords
- Tạo vector cho từ

Tùy thuộc vào tính chất tập dữ liệu mà các bước trên có thể bị lượt bỏ để rút ngắn thời gian xử lý.



Hình 3.2 Tóm tắt các bước tiền xử lý văn bản [23]

Làm sạch văn bản: mục đích bước này là loại bỏ "noise" trong dữ liệu. Đa phần dữ liệu noise là dữ liệu chứa các thẻ HTML, JavaScript. Ví dụ câu "Lập trình NLP" sau khi làm sạch sẽ thu được câu "Lập trình NLP".

Tách từ: trong Tiếng Việt dấu cách (space) chỉ mang ý nghĩa phân cách âm tiết với nhau, vì vậy để xử lý trong Tiếng Việt công đoán tách từ (word segmentation) là một trong những bài toán quan trọng. Ví dụ trong câu "Giảng viên dạy tốt" nếu tách thành 4 từ độc lập thì từ "Giảng" và "viên" sẽ mang ý nghĩa riêng biệt nhau so với khi chúng đứng cùng nhau trong từ "Giảng viên".

Chuẩn hóa từ: mục đích đưa các văn bản không đồng nhất về cùng một dạng. Ví dụ ta sẽ chuẩn hóa các từ viết tắt như "k", "ko", "k0" về đúng chuẩn là "không".

Loại bỏ Stopwords: là loại bỏ những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở Tiếng Việt Stopwords là những từ như: để, là, này, kia,.v.v.

Trong luận văn này tôi đã sử dụng bộ thư viện ViTokenizer được xây dựng và phát triển bởi tác giả Trần Trung Việt, đây là bộ thư viện mã nguồn mở trong python. Bộ thư viện này được sử dụng rất rộng rãi và có độ hiệu quả cao trong việc tiền xử lý văn bản.

Sau khi thực hiện đầy đủ quy trình tiền xử lý tôi chia dữ liệu theo tỷ lệ 90:10 để sử dụng làm dữ liệu train và validation.

3.3.2 Biểu diễn văn bản

Ở luận văn này sau khi tiền xử lý văn bản tôi sẽ biểu diễn văn bản trong mô hình không gian vector. Tôi sẽ sử dụng hai mô hình phổ biến là Word2vec cho việc tạo Pretrained word embedding và Sentence2vec cho việc biểu diễn các dữ liệu ý kiến đánh giá của sinh viên.

3.3.2.1. Xây dựng model pretraind word embedding

Đầu tiên tôi sẽ tiến hành tạo một model pretrained word embedding bằng thư viện gensim. Model này sẽ được biểu diễn bằng mô hình Word2vec trong không gian vector. Nguồn dữ liệu này sử dụng cho việc tạo model được lấy từ bài tổng hợp "**A Large-scale**

Vietnamese News Text Classification Corpus” của nhóm tác giả Cong Duy Vu Hoang, Dien Dinh, Le Nguyen Nguyen, Quoc Hung Ngo vào năm 2016. Chi tiết dữ liệu sử dụng được thể hiện qua hình ảnh bên dưới:

******Train*****

Topic	Topic ID	#files
Am nhac	AN	900
Am thuc	AT	265
Bat dong san	BDS	246
Bong da	BD	1857
Chung khoan	CK	382
Cum ga	CG	510
Cuoc song do day	CSDD	729
Du hoc	DH	682
Du lich	DL	582
Duong vao WTO	DVW	208
Gia dinh	GD	213
Giai tri tin hoc	GTTH	825
Giao duc	GDu	821
Gioi tinh	GT	343
Hacker & virus	HV	355
Hinh su	HS	155
Khong gian song	KGS	134
Kinh doanh quoc te	KDQT	571
Lam dep	LD	776
Loi song	LS	223
Mua sam	MS	187
My thuat	MT	193
San khau dien anh	SKDA	1117
San pham tin hoc moi	SPTHM	770
Tennis	T	588
The gioi tre	TGT	331
Thoi trang	TT	412
Total		14375

Hình 3.3 Tổng hợp danh sách dữ liệu pretrained word embedding

Ở đây tôi sử dụng mô hình Word2vec của thư viện gensim với các thông số kỷ thuật sau:

- Mô hình: Skip-gram (sg=1)

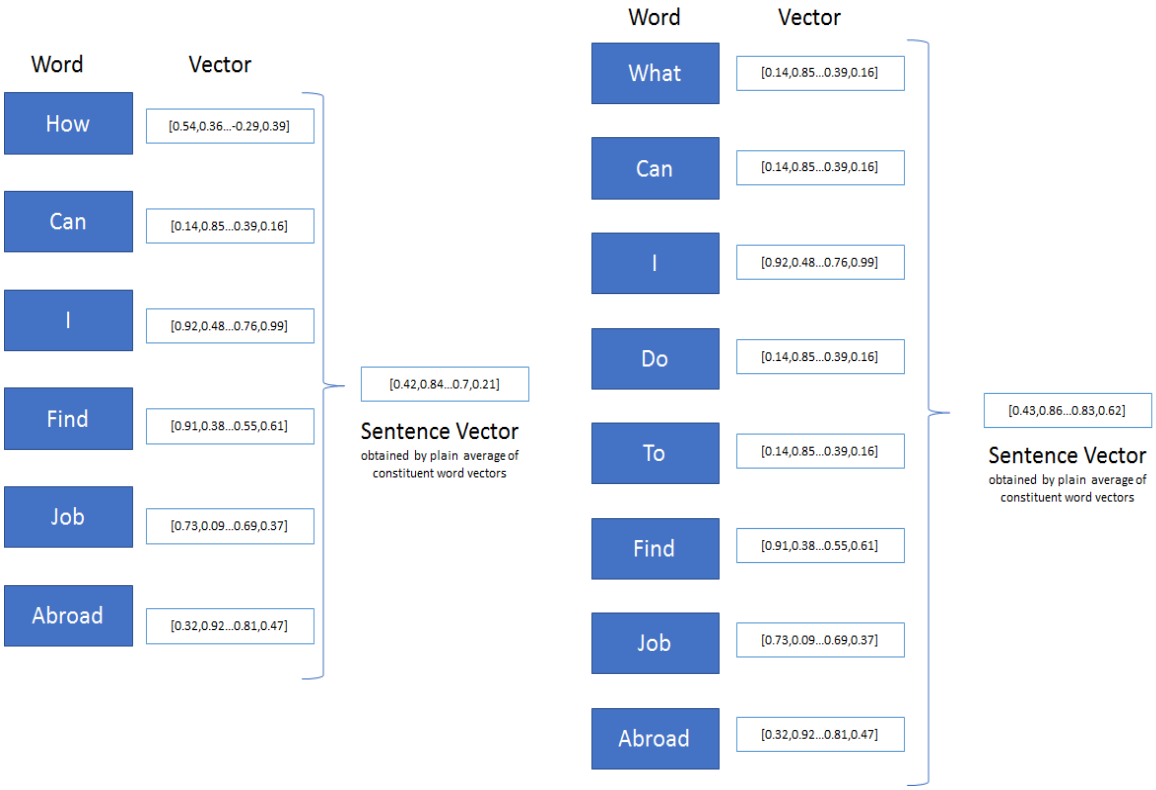
- Số chiều vector: 200
- Độ trượt từ: 5 (window)
- Độ dài tối thiểu: 5 (min count)
- Lần lặp: 10 (inter)

Sau khi xây dựng được mô hình pretrained word embedding tôi sẽ lưu tạm mô hình này vào một tệp model. Model này sẽ được dùng làm cơ sở ánh xạ dữ liệu đánh giá thành mô hình sentence2vec ở bước tiếp theo.

3.3.2.2. Biểu diễn văn bản sang mô hình sentence2vec

Từ tệp model word embedding word2vec đã xử lý được kết hợp với những dữ liệu ý kiến đánh giá đã được tiền xử lý. Ở bước này tôi sẽ thực hiện việc ánh xạ từng câu ý kiến đánh giá thành các vector (sentence2vec) thông qua bộ word embedding.

Việc mô hình hóa các câu dữ liệu sang không gian vector ở đây tôi thực hiện theo cách đơn giản nhất là tính trung bình trọng số vector của các từ trong câu dựa theo nghiên cứu của nhóm tác giả Sanjeev Arora, Yingyu Liang, Tengyu Ma tại [24].



Hình 3.4 Mô hình xây dựng sentence2vec cho câu [24]

Phương pháp sentence2vec này thực hiện rất đơn giản những tồn tại những hạn chế nhất định như:

- Nó bỏ qua thứ tự của các từ trong câu.
- Nó hoàn toàn bỏ qua ngữ nghĩa, ngữ cảnh của câu.

Ví dụ như 2 câu sau có cùng thành phần nhưng mang hai ý nghĩa hoàn toàn trái ngược:

- Vì quá yêu bạn gái, chàng trai quyết định từ bỏ game.
- Vì quá yêu game, chàng trai quyết định từ bỏ bạn gái.

3.3.3 Phân lớp cảm xúc

Phân lớp cảm xúc là bước quan trọng trong bài toán dự đoán cảm xúc đã đặt ra vì vậy để dữ liệu mang tính ngẫu nhiên và khách quan. Tôi sẽ dùng một số kỷ thuật để xốc dữ liệu lên một cách ngẫu nhiên cụ thể ở đây tôi sẽ dùng thư viện KFold trong Python để thực hiện xốc dữ liệu ở mỗi lần chạy thực nghiệm.

Dữ liệu sau khi được xốc sẽ được chia nhỏ thành 10 phần bằng nhau 100 dữ liệu/phần và lấy ngẫu nhiên theo tỷ lệ 9/1, 9 phần sẽ được dùng để làm dữ liệu huấn luyện và 1 phần dùng làm dữ liệu test. Việc xốc dữ liệu và lấy ngẫu nhiên dữ liệu từ các phần chia nhỏ sẽ đảm bảo dữ liệu đưa vào các bộ phân lớp mang tính khách quan trong việc đánh giá.

Ở bước tiếp từ những dữ liệu huấn luyện đã được chọn ra, tôi sẽ lần lượt mô hình hóa các dữ liệu này sang mô hình sentence2vec kèm theo nhãn dán đã gán ở bước tiền xử lý. Như vậy sau bước này ta sẽ có được tập dữ liệu gồm các vector tương ứng với câu kèm theo nhãn gán để sử dụng huấn luyện trong các bộ phân lớp.

$$\begin{Bmatrix} s_1, l_1 \\ s_2, l_2 \\ \dots \\ s_i, l_j \end{Bmatrix}$$

trong đó s_i là 1 vector tương ứng của 1 ý kiến đánh giá, l_i là nhãn tương ứng của vector s_i

Tiếp theo bộ dữ liệu vector kèm theo nhãn dán sẽ được đưa vào huấn luyện trong các bộ phân lớp, ở đây tôi sẽ sử dụng nhiều bộ phân lớp khác nhau để có thể so sánh độ hiệu quả giữa các phương pháp trên cùng tập dữ liệu ý kiến đánh giá này.

Mô hình phân lớp được sử dụng trong luận văn là:

- Mô hình phân lớp Support Vector Machine (SVM)
- Mô hình phân lớp Naïve Bayes

Ở luận văn này để triển khai các bộ phân lớp tôi sử dụng một thư viện khá phổ viện trong Python là Sklearn, thư viện này hỗ trợ nhiều phương pháp phân lớp khác nhau phù hợp với quy mô luận văn. Chi tiết quá trình quá trình phân lớp sẽ được mô tả ở phần thực nghiệm.

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, Pages 3788-3792, 2010.

[5] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool, 2012

[6] B. Pang & L. Lee & S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP, Page 79-86, 2002.

[7] 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.

[8] G. Qiu & B. Liu & J. Bu & C. Chen, Opinion word expansion and target extraction through double Propagation, Journal Computational Linguistics, Page 9-27, 2011.

[9] 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.

[10] 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.

[11] Stanford University (2019). Text Classification and Naïve Bayes [online], viewed 12 March 2019, from:< https://web.stanford.edu/class/cs124/lec/naivebayes.pdf>

[12] V.N. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[13] 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

[14] T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, 2005

[15] Ratnaparkhi, A. (1997), "A Simple Introduction to Maximum Entropy Models for Natural Language Processing", IRCS Technical Reports Series.

[16] E. Riloff & J. Wiebe, Learning Extraction Patterns for Subjective Expressions, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Page 105-112, 2003.

[17] Đặng. Thái & Lê. Cường & Nguyễn. Hương & Huỳnh. Nam, Automatically Learning Patterns in Subjectivity Classification for Vietnamese, Advances in Intelligent Systems and Computing, 2015

[18] Nguyễn. Anh, "Nghiên cứu kỹ thuật đánh giá độ tương đồng văn bản ứng dụng so sánh văn bản tiếng Việt", Đại học Hàng Hải, 2016.

[19] J. Sowa, "Conceptual Graphs For Representing Conceptual Structures", 2009

[20] Nguyễn Văn Quý, "Nghiên cứu các phương pháp chuẩn hóa chữ viết tắt trong văn bản tiếng Việt", Đại học Bách khoa, Đại học Đà Nẵng, 2017

[21] Mikolov & Tomas, "Efficient Estimation of Word Representations in Vector Space", 2013.

[22] "Bàn về công đoạn tiền xử lý trong xử lý ngôn ngữ tự nhiên" Bayes [online],viewed 12 March 2019, from:< https://blog.vietnamlab.vn/2018/01/24/ban-ve-cong-doan-tien-xu-ly-trong-xu-ly-ngon-ngu-tu-nhien>

[23] C. Angermueller & T. Pärnamaa & L. Parts & O. Stegle, "Deep Learning for Computational Biology", 2016

[24] Sanjeev Arora, Yingyu Liang, Tengyu Ma, "A SIMPLE BUT TOUGH TO BEAT BASELINE FOR SENTENCE EMBEDDINGS" [online], viewed 12 March 2019, from:< https://github.com/peter3125/sentence2vec>

[25]