ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

KHỔNG BÙI TRUNG

PHÂN LOẠI GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI DỰA VÀO TIN NHẮN VĂN BẢN VÀ WORD2VEC

LUẬN VĂN THẠC SĨ KỸ THUẬT PHẦN MỀM

Hà Nội – 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

KHỔNG BÙI TRUNG

PHÂN LOẠI GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI DỰA VÀO TIN NHẮN VĂN BẢN VÀ WORD2VEC

Ngành: Công nghệ thông tin
Chuyên ngành: Kỹ thuật phần mềm
Mã số: 60480103

LUẬN VĂN THẠC SĨ KỸ THUẬT PHẦN MỀM

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. NGUYỄN VĂN VINH

Hà Nội – Năm 2016

LỜI CÁM ƠN

Để có được kết quả như ngày hôm nay, tôi luôn ghi nhớ công ơn của các thầy cô, bạn bè, đồng nghiệp và gia đình, những người đã dạy bảo và ủng hộ tôi trong suốt quá trình học tập.

Trước hết, tôi muốn gửi lời cám ơn đến các thầy cô trường Đại học Công Nghê, Đại học Quốc Gia Hà Nội đã quan tâm tổ chức chỉ đạo và trực tiếp giảng dạy khoá cao học của tôi. Đặc biệt, tôi xin gửi lời cảm ơn sâu sắc đến thầy giáo hướng dẫn TS Nguyễn Văn Vinh, người đã tận tình chỉ bảo và góp ý về mặt chuyên môn cho tôi trong suốt quá trình làm luận văn. Nếu không có sự giúp đỡ của thầy thì tôi khó có thể hoàn thành được luận văn này.

Cũng qua đây, tôi xin gửi lời cảm ơn đến ban lãnh đạo Trường TCN Nấu ăn và NVKS Hà Nội, nơi tôi công tác, đã tạo mọi điều kiện thuận lợi cho tôi trong thời gian hoàn thành các môn học cũng như trong suốt quá trình làm luận văn tốt nghiệp.

Cuối cùng, tôi xin cảm ơn gia đình và các bạn bè, đồng nghiệp đã luôn ủng hộ, động viên để tôi yên tâm nghiên cứu và hoàn thành luận văn.

Trong suốt quá trình làm luận văn, bản thân tôi đã cố gắng tập trung tìm hiểu, nghiên cứu và tham khảo thêm nhiều tài liệu liên quan. Tuy nhiên, do bản thân mới bắt đầu trên con đường nghiên cứu khoa học, chắc chắn bản luận văn vẫn còn nhiều thiếu sót. Tôi rất mong được nhận sự chỉ bảo của các Thầy Cô giáo và các góp ý của bạn bè đồng nghiệp để luận văn được hoàn thiện hơn.

Hà Nội, Tháng 11 năm 2016

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

(Ký và ghi rõ họ tên)

Khổng Bùi Trung

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

Ngày nay, con người đang sở hữu kho dữ liệu phong phú, đa dạng và khổng lồ. Đặc biệt sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực đã làm cho kho dữ liệu ấy tăng lên nhanh chóng. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Mặt khác, trong môi trường cạnh tranh thì người ta ngày càng cần có thông tin với tốc độ nhanh chóng để giúp cho việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên khối lượng dữ liệu khổng lồ đã có. Tiến hành các công việc như vậy chính là quá trình phát hiện tri thức trong cơ sở dữ liệu, trong đó kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền Công nghệ thông tin thế giới hiện nay nói chung và Việt Nam nói riêng. Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất kinh doanh của mình và thu được những lợi ích to lớn.

Hiện nay mạng xã hội như Facebook, Twitter, Zalo,… ngày càng phát triển và có ảnh hưởng lớn đến đời sống xã hội. Trong lĩnh vực thương mại điện tử, nhiều công ty vào mạng xã hội để quảng cáo, tư vấn, phân tích về sản phẩm và công ty của mình. Chính vì vậy nếu biết được giới tính người dùng là nam hay nữ thì việc tư vấn và quảng cáo hướng đến người dùng sẽ cụ thể và hiệu quả hơn.

Do đó vấn đề phân loại tự động giới tính của người dùng sử dụng mạng xã hội là một bài toán quan trọng. Hiện nay có rất nhiều kỹ thuật để sử dụng cho phân loại tự động giới tính nhưng chủ yếu là dựa vào các đặc trưng kiểu truyền thống như trong mô hình tần suất từ, n-gram,... Word2Vec và mô hình chuyển từ thành vector được phát triển và ứng dụng rộng rãi trong thời gian gần đây. Chính vì vậy mà chúng tôi sử dụng thêm Word2Vec làm đặc trưng để cải tiến kết quả bài toán này.

Từ những vấn đề nêu trên, chúng tôi chọn đề tài: ***"Phân loại giới tính người dùng mạng xã hội dựa trên tin nhắn văn bản và Word2Vec"*** để làm luận văn tốt nghiệp.

Đề tài này nhằm mục đích nghiên cứu phương pháp biểu diễn các từ dưới dạng vector sau đó dùng làm đặc trưng để cải thiện kết quả của việc phân loại giới tính người dùng mạng xã hội dựa vào tin nhắn văn bản.

Luận văn bao gồm phần Mở đầu, phần kết luận và ba chương.

Phần mở đầu sẽ giới thiệu về đề tài luận văn. Phần này sẽ trình bày lý do của đề tài, mục tiêu của đề tài và cấu trúc của luận văn.

Chương 1 giới thiệu tổng quan về khai phá dữ liệu và quá trình khai phá dữ liệu. Bên cạnh đó còn giới thiệu một số chức năng chính của khai phá dữ liệu cũng như một số kỹ thuật khai phá dữ liệu. Ngoài ra chương này còn giới thiệu về mạng xã hội, các lợi ích và bất lợi của mạng xã hội cũng như một số mạng xã hội phổ biến trên thế giới hiện nay.

Chương 2 giới thiệu khái niệm về vector từ cũng như các lập luận liên quan đến vector từ. Chương này còn giới thiệu về các mô hình cũng như cách xây dựng một Word2Vec như mô hình Continuous Bag-of-Words, mô hình Skip-gram.

Chương 3 trình bày về về thực nghiệm bài toán ứng dụng Word2Vec vào phân loại giới tính người dùng mạng xã hội. Giải pháp thực hiện và các kết quả đạt được sau khi thực nghiệm.

Cuối cùng là phần kết luận, định hướng nghiên cứu phát triển đề tài và những tài liệu tham khảo của luận văn.

CHƯƠNG 1: TỔNG QUAN KHAI PHÁ DỮ LIỆU VÀ MẠNG XÃ HỘI

1.1. Khai phá dữ liệu

1.1.1. Khai phá dữ liệu là gì?

Khai phá dữ liệu (datamining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Một ví dụ hay được sử dụng là việc khai thác vàng từ đá và cát, Dataming được ví như công việc "Đãi cát tìm vàng" trong một tập hợp lớn các dữ liệu cho trước. Thuật ngữ Datamining ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction (chắt lọc tri thức), data/patern analysis (phân tích dữ liệu/mẫu), data archaeoloogy (khảo cổ dữ liệu), datadredging (nạo vét dữ liệu), ... [1].

Sau đây là một số định nghĩa mang tính mô tả của nhiều tác giả về khai phá dữ liệu:

Định nghĩa của Ferruzza: “Khai phá dữ liệu là tập hợp các phương pháp được dùng trong tiến trình khám phá tri thức để chỉ ra sự khác biệt các mối quan hệ và các mẫu chưa biết bên trong dữ liệu”.

Định nghĩa của Parsaye: “Khai phá dữ liệu là quá trình trợ giúp quyết định, trong đó ta tìm kiếm các mẫu thông tin chưa biết và bất ngờ trong CSDL lớn”.

Định nghĩa của Fayyad: “Khai phá tri thức là một quá trình không tầm thường nhận ra những mẫu dữ liệu có giá trị, mới, hữu ích, tiềm năng và có thể hiểu được”.

1.1.2. Quá trình khai phá dữ liệu

Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau [1]:

1. Làm sạch dữ liệu (data cleaning & preprocessing): Loại bỏ nhiễu và các dữ liệu không cần thiết.

2. Tích hợp dữ liệu: (data integration): quá trình hợp nhất dữ liệu thành những kho dữ liệu (data warehouses & data marts) sau khi đã làm sạch và tiền xử lý (data cleaning & preprocessing).

3. Trích chọn dữ liệu (data selection): trích chọn dữ liệu từ những kho dữ liệu và sau đó chuyển đổi về dạng thích hợp cho quá trình khai thác tri thức. Quá trình này bao gồm cả việc xử lý với dữ liệu nhiễu (noisy data), dữ liệu không đầy đủ (incomplete data), ...

4. Chuyển đổi dữ liệu: Các dữ liệu được chuyển đổi sang các dạng phù hợp cho quá trình xử lý.

5. Khai phá dữ liệu (data mining): Là một trong các bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để chắt lọc ra những mẫu dữ liệu.

6. Ước lượng mẫu (knowledge evaluation): Quá trình đánh giá các kết quả tìm được thông qua các độ đo nào đó.

7. Biểu diễn tri thức (knowledge presentation): Quá trình này sử dụng các kỹ thuật để biểu diễn và thể hiện trực quan cho người dùng.

1.1.3. Các chức năng chính của khai phá dữ liệu

Data Mining được chia nhỏ thành một số hướng chính như sau [1]:

• Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.

• Luật kết hợp (association rules): là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Ví dụ: "60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, ...

• Phân loại và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân loại vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của "machine learning" như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), ... Người ta còn gọi phân loại là học có giám sát (học có thầy).

• Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).

• Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

1.1.4. Các kỹ thuật khai phá dữ liệu

1.1.4.1. Phân loại (phân loại - classification)

Là việc xác định một hàm ánh xạ từ một mẫu dữ liệu vào một trong số các lớp đã được biết trước đó. Mục tiêu của thuật toán phân loại là tìm ra mối quan hệ nào đó giữa thuộc tính dự báo và thuộc tính phân loại. Như thế quá trình phân loại có thể sử dụng mối quan hệ này để dự báo cho các mục mới. Các kiến thức được phát hiện biểu diễn dưới dạng các luật theo cách sau: "Nếu các thuộc tính dự báo của một mục thoả mãn điều kiện của các tiền đề thì mục nằm trong lớp chỉ ra trong kết luận" [3].

Ví dụ: Một mục biểu diễn thông tin về nhân viên có các thuộc tính dự báo là: họ tên, tuổi, giới tính, trình độ học vấn, … và thuộc tính phân loại là trình độ lãnh đạo của nhân viên.

1.1.4.2. Hồi qui (regression)

Là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân loại, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không phải rời rạc. Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển, chẳng hạn như hồi quy tuyến tính. Tuy nhiên, phương pháp mô hình hoá cũng được sử dụng, ví dụ: cây quyết định.

Ứng dụng của hồi quy là rất nhiều, ví dụ: dự đoán số lượng sinh vật phát quang hiện thời trong khu rừng bằng cách dò tìm vi sóng bằng các thiết bị cảm biến từ xa; ước lượng sác xuất người bệnh có thể chết bằng cách kiểm tra các triệu chứng; dự báo nhu cầu của người dùng đối với một sản phẩm, … [3].

1.1.4.3. Phân cụm (clustering)

Là việc mô tả chung để tìm ra các tập hay các nhóm, loại mô tả dữ liệu. Các nhóm có thể tách nhau hoặc phân cấp hay gối lên nhau. Có nghĩa là dữ liệu có thể vừa thuộc nhóm này lại vừa thuộc nhóm khác. Các ứng dụng khai phá dữ liệu có nhiệm vụ phân nhóm như phát hiện tập các khách hàng có phản ứng giống nhau trong CSDL tiếp thị; xác định các quang phổ từ các phương pháp đo tia hồng ngoại, … Liên quan chặt chẽ đến việc phân nhóm là nhiệm vụ đánh giá dữ liệu, hàm mật độ xác suất đa biến/các trường trong CSDL [3].

1.1.4.4. Tổng hợp (summarization)

Là công việc liên quan đến các phương pháp tìm kiếm một mô tả tập con dữ liệu. Kỹ thuật tổng hợp thường áp dụng trong việc phân tích dữ liệu có tính thăm dò và báo cáo tự động. Nhiệm vụ chính là sản sinh ra các mô tả đặc trưng cho một lớp. Mô tả loại này là một kiểu tổng hợp, tóm tắt các đặc tính chung của tất cả hay hầu hết các mục của một lớp. Các mô tả đặc trưng thể hiện theo luật có dạng sau: "Nếu một mục thuộc về lớp đã chỉ trong tiền đề thì mục đó có tất cả các thuộc tính đã nêu trong kết luận". Lưu ý rằng luật dạng này có các khác biệt so với luật phân loại. Luật phát hiện đặc trưng cho lớp chỉ sản sinh khi các mục đã thuộc về lớp đó [3].

1.1.4.5. Mô hình hoá sự phụ thuộc (dependency modeling)

Là việc tìm kiếm một mô hình mô tả sự phụ thuộc giữa các biến, thuộc tính theo hai mức: Mức *cấu trúc của mô hình* mô tả (thường dưới dạng đồ thị). Trong đó, các biến phụ thuộc bộ phận vào các biến khác. Mức *định lượng mô hình* mô tả mức độ phụ thuộc. Những phụ thuộc này thường được biểu thị dưới dạng theo luật "*nếu - thì*" (nếu tiền đề là đúng thì kết luận đúng). Về nguyên tắc, cả tiền đề và kết luận đều có thể là sự kết hợp logic của các giá trị thuộc tính. Trên thực tế, tiền đề thường là nhóm các giá trị thuộc tính và kết luận chỉ là một thuộc tính. Hơn nữa hệ thống có thể phát hiện các luật phân loại trong đó tất cả các luật cần phải có cùng một thuộc tính do người dùng chỉ ra trong kết luận.

Quan hệ phụ thuộc cũng có thể biểu diễn dưới dạng mạng tin cậy Bayes. Đó là đồ thị có hướng, không chu trình. Các nút biểu diễn thuộc tính và trọng số của liên kết phụ thuộc giữa các nút đó [3].

1.1.4.6. Phát hiện sự biến đổi và độ lệch (change and deviation dectection)

Nhiệm vụ này tập trung vào khám phá hầu hết sự thay đổi có nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, phát hiện độ lệch đáng kể giữa nội dung của tập con dữ liệu thực và nội dung mong đợi. Hai mô hình độ lệch hay dùng là *lệch theo thời gian* hay *lệch theo nhóm*. Độ lệch theo thời gian là sự thay đổi có ý nghĩa của dữ liệu theo thời gian. Độ lệch theo nhóm là sự khác nhau giữa dữ liệu trong hai tập con dữ liệu, ở đây tính cả trường hợp tập con dữ liệu này thuộc tập con kia, nghĩa xác định dữ liệu trong một nhóm con của đối tượng có khác đáng kể so với toàn bộ đối tượng không? Theo cách này, sai sót dữ liệu hay sai lệch so với giá trị thông thường được phát hiện. Vì những nhiệm vụ này yêu cầu số lượng và các dạng thông tin rất khác nhau nên thường ảnh hưởng đến việc thiết kế và chọn phương pháp khai phá dữ liệu khác nhau [3].

1.2. Mạng xã hội

1.2.1. Mạng xã hội là gì?

Mạng xã hội là việc thực hiện mở rộng một số lượng các mối quan hệ của doanh nghiệp hoặc các quan hệ xã hội bằng cách tạo ra các kết nối thông qua các cá nhân người dùng, thường là thông qua các trang web mạng xã hội như Facebook, Twitter, LinkedIn và Google+[16].

Dựa trên sáu cấp độ của khái niệm ngăn cách (ý tưởng rằng bất kỳ hai người trên hành tinh này có thể thực hiện liên lạc thông qua một chuỗi không quá năm người trung gian), mạng xã hội thiết lập các cộng đồng trực tuyến kết nối với nhau (đôi khi được gọi là đồ thị xã hội) giúp con người liên lạc được với những người họ biết – những người họ không thể gặp bằng phương thức khác [16].

Tùy thuộc vào các nền tảng truyền thông xã hội, các thành viên có thể liên hệ với bất kỳ thành viên khác. Trong trường hợp khác, các thành viên có thể liên hệ với bất cứ ai họ có một kết nối đến, và sau đó là bất cứ ai liên lạc có một kết nối đến, và cứ như vậy. Một số dịch vụ yêu cầu các thành viên phải có một kết nối từ trước để liên hệ với các thành viên khác [16].

Trong khi mạng xã hội đã đi vào gần như là hầu hết các lĩnh vực đang tồn tại ở xã hội, với tiềm năng vô cùng lớn của Web, để tạo điều kiện kết nối như vậy đã dẫn đến việc mở rộng theo cấp số nhân và liên tục của mạng xã hội.

Ngoài nền tảng truyền thông xã hội, khả năng tương tác xã hội và hợp tác ngày càng được xây dựng và mở rộng vào các ứng dụng kinh doanh [16].

1.2.2. Lợi ích và tác hại của mạng xã hội

1.2.2.1. Lợi ích của mạng xã hội

Mạng xã hội ngày nay có một số lợi ích như sau [4]:

a. Giới thiệu bản thân mình với mọi người: ta có thể giới thiệu tính cách, sở thích, quan điểm của bản thân trên mạng xã hội và nó có thể giúp ta tìm kiếm những cơ hội phát triển khả năng của bản thân.

b. Kết nối bạn bè: ta có thể biết được nhiều thông tin về bạn bè hoặc người thân bằng cách kết bạn trên mạng xã hội. Ta cũng có thể gặp gỡ và giao lưu kết bạn với tất cả mọi người trên thế giới có cùng sở thích hay quan điểm giống mình. Từ đó có thể xây dựng mối quan hệ tốt đẹp hơn hoặc hợp tác với nhau về nhiều mặt.

c. Tiếp nhận thông tin, học hỏi kiến thức và kỹ năng: việc cập nhật thông tin trong một xã hội hiện đại như hiện nay là điều nên làm và cần phải làm, nó giúp ta dễ dàng tìm hiểu, nắm bắt được nhiều thông tin quan trọng. Học hỏi thêm rất nhiều kiến thức, trau dồi những kĩ năng giúp cho bạn hoàn thiện bản thân mình hơn nữa.

d. Kinh doanh: bán và mua hàng online không còn xa lạ với tất cả ta vì thế mạng xã hội là một môi trường kinh doanh vô cùng lí tưởng. Ta cũng có thể dùng nó để quảng cáo cho những sản phẩm của công ty, giúp cho ta có thể tìm kiếm được những khách hàng tiềm năng.

e. Bày tỏ quan niệm cá nhân: trải qua rất nhiều hoạt động căng thẳng trong cuộc sống, mỗi con người cần bày tỏ và cần nhận được sự sẻ chia để ta

cảm thấy thanh thản hơn. Thế nhưng việc chia sẻ vấn đề của mình ngoài đời thực đôi khi trở nên khó khăn với một số người ít nói. Chính vì thế việc viết ra những suy nghĩ của mình qua bàn phím máy tính sẽ giúp ta giải tỏa được phần nào.

f. Mang đến lợi ích về sức khoẻ: giúp cải thiện não bộ và làm chậm quá trình lão hoá, nghiên cứu của giáo sư Gary Small tại trường Đại học California Los Angeles cho thấy càng sử dụng và tìm kiếm nhiều thông tin với internet, não bộ sẽ càng được rèn luyện tốt hơn và các khả năng phán đoán, quyết định cũng sẽ từ đó phát triển thêm. Ông còn đồng thời nhận thấy rằng, việc sử dụng internet nhiều có thể giúp cho não bộ hoạt động tốt hơn, giúp làm giảm quá trình lão hóa và làm cho người lớn tuổi vẫn có suy nghĩ hết sức lạc quan.

1.2.2.2. Tác hại của mạng xã hội

Ta không thể phủ nhận những lợi ích mà mạng xã hội đã mang đến cho con người hiện nay như giúp ích cho công việc, cho việc tìm kiếm thông tin, thiết lập các mối quan hệ cá nhân hay giải trí… Tuy nhiên, nó cũng chứa đựng nhiều nguy cơ, rủi ro tiềm ẩn có thể ảnh hưởng xấu tới công việc, mối quan hệ cá nhân và cuộc sống của người sử dụng [4]:

a. Giảm tương tác giữa người với người: nghiện mạng xã hội không chỉ khiến bạn dành ít thời gian cho người thật việc thật ở quanh mình, mà còn khiến họ buồn phiền khi bạn coi trọng bạn bè "ảo" từ những mối quan hệ ảo hơn

những gì ở trước mắt. Dần dần, các mối quan hệ sẽ bị rạn nứt và sẽ chẳng ai còn muốn gặp mặt bạn nữa.

b. Lãng phí thời gian và xao lãng mục tiêu thực của cá nhân: quá chú tâm vào mạng xã hội dễ dàng làm người ta quên đi mục tiêu thực sự của cuộc sống. Thay vì chú tâm tìm kiếm công việc trong tương lai bằng cách học hỏi những kỹ năng cần thiết, các bạn trẻ lại chỉ chăm chú để trở thành *"anh hùng bàn phím"* và nổi tiếng trên mạng. Ngoài ra, việc đăng tải những thông tin *"giật*

gân" nhằm câu like không còn là chuyện xa lạ, song nó thực sự khiến người khác phát bực nếu dùng quá thường xuyên. Mạng xã hội cũng góp phần tăng sự ganh đua, sự cạnh tranh không ngừng nghỉ để tìm like và nó sẽ cướp đi đáng kể quỹ thời gian của bạn.

c. Nguy cơ mắc bệnh trầm cảm: các nghiên cứu gần đây cho thấy những ai sử dụng mạng xã hội càng nhiều thì càng cảm thấy tiêu cực hơn, thậm chí có thể dẫn đến trầm cảm. Điều này đặc biệt nguy hiểm với những ai đã được chẩn đoán mắc bệnh trầm cảm từ trước. Vì thế, nếu bạn phát hiện mình thường xuyên cảm thấy mất tinh thần, có lẽ đã đến lúc tạm biệt "facebook" trong một thời gian.

d. Giết chết sự sáng tạo: mạng xã hội là phương tiện hiệu quả nhất để làm tê liệt và giết chết quá trình sáng tạo. Quá trình lướt các trang mạng xã hội có tác động làm tê liệt não bộ tương tự như khi xem tivi trong vô thức. Nếu hôm nay bạn có kế hoạch làm việc thì hãy tuyệt đối tránh xa các trang mạng xã hội.

e. Không trung thực và bạo lực trên mạng: *"Anh hùng bàn phím"* là một từ không còn xa lạ trong thời gian gần đây. Người ta cảm thấy thoải mái trên mạng nên họ thường nói những điều mà ngoài đời không dám phát biểu hoặc không có thực. Đồng thời vấn nạn bạo lực trên mạng càng nhức nhối thì ngoài đời con người cũng dần trở nên bạo lực hơn hẳn.

f. Thường xuyên so sánh bản thân với người khác: những gì người ta khoe khoang trên mạng không hẳn là con người thật của họ, và việc thường xuyên so sánh những thành tựu của mình với bạn bè trên mạng sẽ ảnh hưởng rất tiêu cực đến tinh thần của bạn. Hãy dừng việc so sánh và nhớ rằng ai cũng có điểm mạnh, điểm yếu của riêng mình. Từ những hành động thực tế để có thể làm tăng giá trị của bản thân là điều cần thiết đối với mỗi ta.

g. Mất ngủ: ánh sáng nhân tạo tỏa ra từ màn hình các thiết bị điện tử sẽ đánh lừa não của bạn làm bạn khó ngủ hơn. Ngoài ra, nhiều bạn trẻ hiện nay sẵn sàng thức thâu đêm chỉ vì đam mê các trò chơi trực tuyến. Thiếu ngủ dẫn đến nhiều hệ lụy nghiêm trọng cho sức khỏe và tinh thần.

h. Thiếu riêng tư: đã có nhiều thông tin cho rằng các trang mạng xã hội bán thông tin cá nhân của người sử dụng, lại thêm nhiều nguy cơ từ hacker,

virus. Những điều này đều cảnh báo rằng sự riêng tư cá nhân đang dần mất đi trong khi mạng xã hội càng phát triển.

Từ việc đó, ta thấy rằng, những thông tin được báo chí đăng hay được truyền tải từ mạng xã hội đã được lan tỏa rộng rãi và được dư luận hết sức quan tâm, mặc dù người đọc hay chia sẻ thông tin đó trên mạng xã hội, đều chưa biết thực hư sự chính xác của thông tin đó ra sao. Xét về góc độ này, ta có thể thấy được mặt trái của mạng xã hội, mọi người đều có thể đọc và chia sẻ những thông tin mà không hiểu rõ về vấn đề, chính điều này đã vô tình gây ra những rắc rối, những ảnh hưởng xấu tới cuộc sống cá nhân của những người trong cuộc.

1.2.3. Các mạng xã hội phổ biến

1.2.3.1. Facebook[1]

Trang mạng xã hội lớn nhất mà ta phải kể đến đó là Facebook. Facebook được xem là mạng xã hội phổ biến và "khủng" nhất trên thế giới ảo với 1,55 tỷ người dùng. Facebook ra đời vào tháng 2 năm 2004 bởi Mark Zuckerburg. Facebook là loại hình mạng xã hội chia sẻ hình ảnh, video, tin nhắn, Blog, v.v… ngoài ra nó còn có ứng dụng nhắn tin nổi tiếng trên Mobile là Whatapp, tích hợp trên hệ điều hành Android, iOS, Windows. Facebook có những ưu điểm mà khiến nhiều người dùng yêu thích sử dụng đó là tích hợp đa ngôn ngữ giúp mọi

[1] https://www.facebook.com/

người trên thế giới dù có khác biệt về ngôn ngữ hay địa lý đều có thể kết nối và tìm thấy được nhau.

1.2.3.2. Instagram[2]

Instagram là một ứng dụng chia sẻ ảnh và video miễn phí trên Apple iOS, Android và Windows Phone. Mọi người có thể tải ảnh hoặc video lên dịch vụ của mình và chia sẻ với người theo dõi của mình hoặc với một nhóm bạn bè chọn lọc. Instagram có 400 triệu người dùng

1.2.3.3. Twitter[3]

Twitter là một trang mạng xã hội cho người sử dụng có thể tải hình ảnh lên, viết và đọc nội dung có độ dài giới hạn. Nếu như bạn là người chuyên nhắn tin điện thoại thì bạn sẽ biết rõ giới hạn 160 ký tự của tin nhắn SMS. Twitter cũng gần giống thế nhưng thậm chí số ký tự cho phép còn ít hơn chỉ có 140 ký tự. Twitter có 320 triệu người dùng.

1.2.3.4. Zalo[4]

Phần mềm Zalo là ứng dụng nhắn tin và gọi điện miễn phí hoạt động trên nền tảng di động. Ưu điểm phần mềm zalo là một ứng dụng cho phép người dùng trò chuyện, nhắn tin, gọi điện miễn phí. Ngoài ra, zalo còn là một mạng xã hội thân thiện với người dùng Việt Nam, đặc biệt là giới trẻ. Lần đầu tiên, người Việt đã phát triển được một mạng xã hội có người dùng rộng rãi, phổ biến. Zalo được phát triển bởi tập đoàn game vng – một tập đoàn game của người Việt. Vì vậy, từ giao diện đến từ ngữ, các chức năng đều rất sát với cuộc sống hàng ngày, đều gắn liền với văn hóa ngôn ngữ Việt. Chính vì lẽ đó mà zalo rất dễ sử dụng. Nhiều mạng xã hội nước ngoài rất hay nhưng để sử dụng được nó, đó là cả một vấn đề.

---

[2] https://www.instagram.com/

[3] https://twitter.com

[4] http://zalo.me/

CHƯƠNG 2: WORD2VEC VÀ MÔ HÌNH "TỪ" THÀNH "VECTOR"

2.1. Vector từ là gì

Để máy tính có thể hiểu được các từ thì chúng ta phải biểu diễn các từ đó dưới dạng vector từ. Vector từ là một vector của các trọng số biểu diễn cho từ. Trong dạng biểu diễn 1-of-N (hay "one-hot") việc mã hóa các thành phần trong vector được liên kết với một từ trong bộ từ vựng. Việc mã hóa một từ cho trước là đưa ra một vector, trong đó các phần tử liên quan được thiết lập giá trị là 1, tất cả các phần tử khác là 0.

Giả sử bộ từ vựng của ta chỉ có 5 từ: Vua, Hoàng hậu, Đàn ông, Phụ nữ và Trẻ con. Ta sẽ mã hóa cho từ Hoàng hậu như sau:

0	**1**	**0**	**0**	**0**
Vua	Hoàng hậu	Đàn ông	Phụ nữ	Trẻ con

Hình 2.1: Mã hóa 1-of-N

Trong Word2Vec, một biểu diễn phân tán của một từ được sử dụng. Tạo ra một vector với kích thước vài trăm chiều. Mỗi từ được biểu diễn bởi tập các trọng số của từng phần tử trong nó. Vì vậy, thay vì sự kết nối 1-1 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.

Nếu ta gán nhãn các kích thước cho một vector từ giả thuyết, nó trông giống như hình sau:

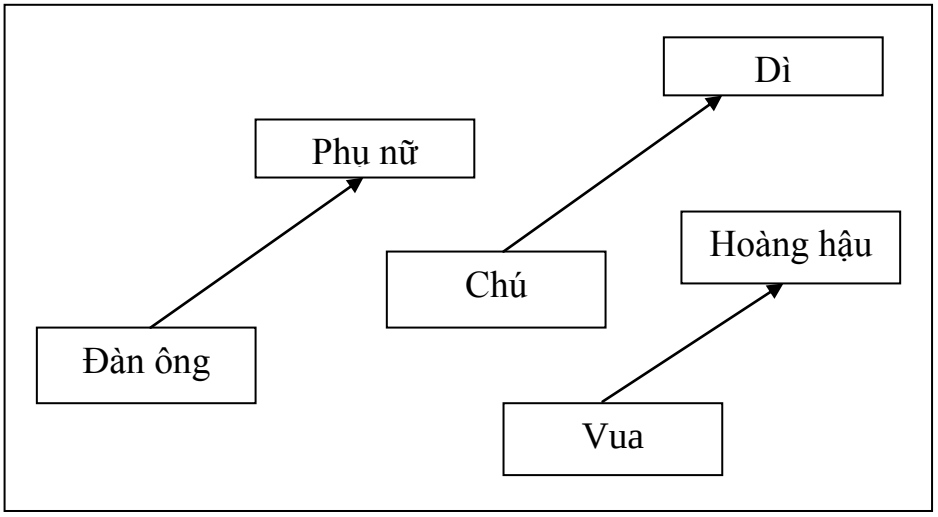
		Vua	Hoàng hậu	Phụ nữ	Công chúa
Hoàng gia		0.99	0.99	0.02	0.98
Nam tính		0.99	0.05	0.01	0.02
Nữ tính		0.05	0.93	0.999	0.94
Tuổi		0.7	0.6	0.5	0.1

Như vậy một vector trở thành đại diện một cách tóm lược ý nghĩa của một từ. Và như ta sẽ thấy tiếp theo, đơn giản bằng việc kiểm tra một tập văn bản lớn, nó có thể học các vector từ, ta có thể nắm bắt mối quan hệ giữa các từ theo một cách đáng ngạc nhiên. Ta cũng có thể sử dụng các vector như các đầu vào cho một mạng Nerual.

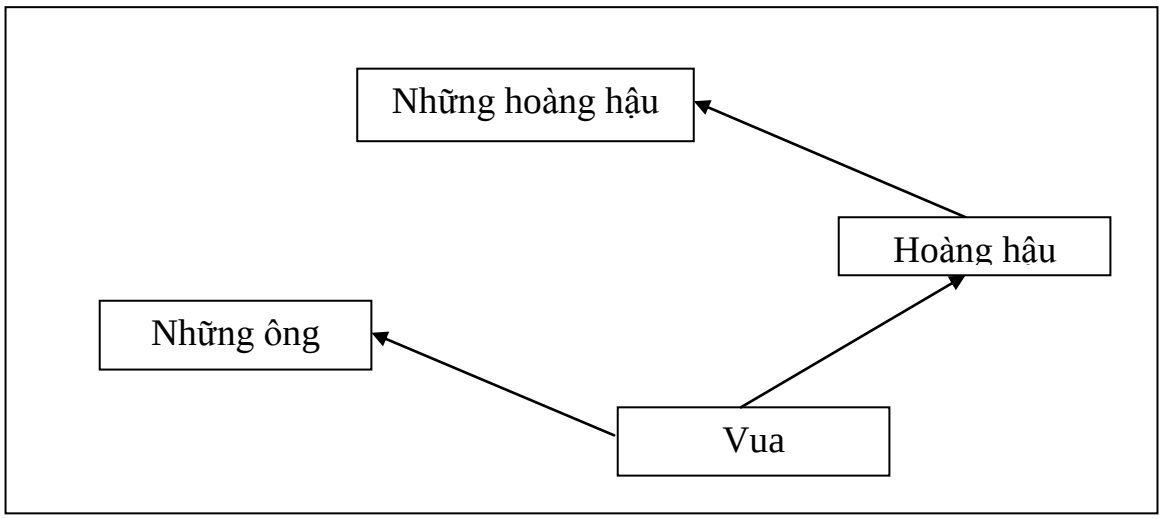
2.2. Lập luận với Vector từ

Ta thấy rằng các đại diện từ được nghiên cứu trong thực tế nắm bắt quy tắc cú pháp và ngữ nghĩa có ý nghĩa theo một cách rất đơn giản. Cụ thể, các quy tắc được quan sát như các giá trị bù vector không đổi giữa các cặp từ chia sẻ một mối quan hệ đặc biệt. Ví dụ, nếu ta ký hiệu vector cho chữ i là X_i, và tập trung vào mối quan hệ số ít/số nhiều, ta sẽ quan sát thấy rằng $X_{apple} - X_{apples} \approx X_{car} - X_{cars}$, $X_{family} - X_{families} \approx X_{car} - X_{cars}$, v.v. Ta thấy rằng đây cũng là trường hợp cho một loạt các quan hệ ngữ nghĩa được đo bởi mối quan hệ tương đồng [7].

Các vector rất tốt khi trả lời câu hỏi tương tự dạng a là dành cho b như c là dành cho?. Ví dụ, Man (đàn ông) là dành cho Woman (phụ nữ) như uncle (chú) là dành cho? Aunt (thím, dì) sử dụng một phương pháp các giá trị bù vector đơn giản dựa vào khoảng cách cosin.



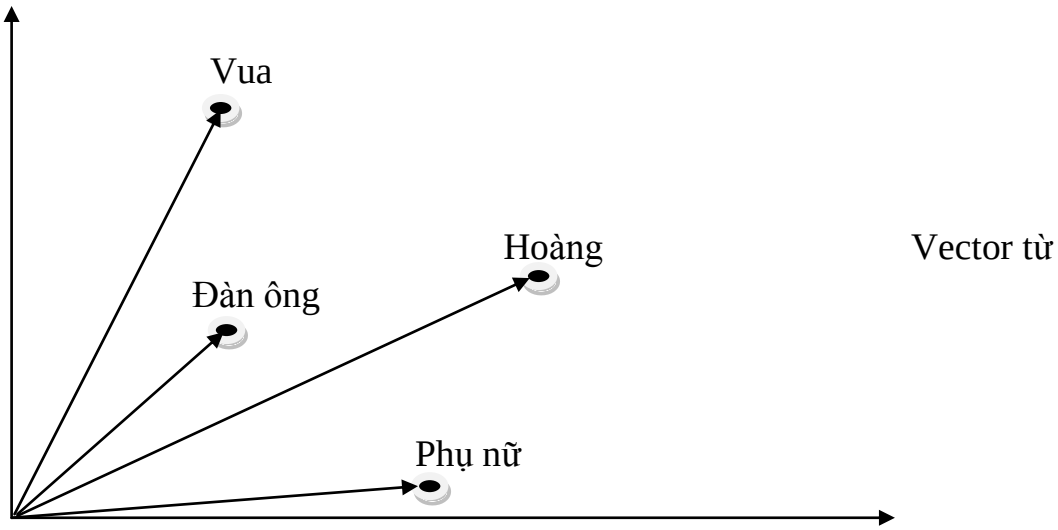
Hình 2.1: Giá trị bù vector cho 3 cặp từ mô phỏng mối quan hệ về giới



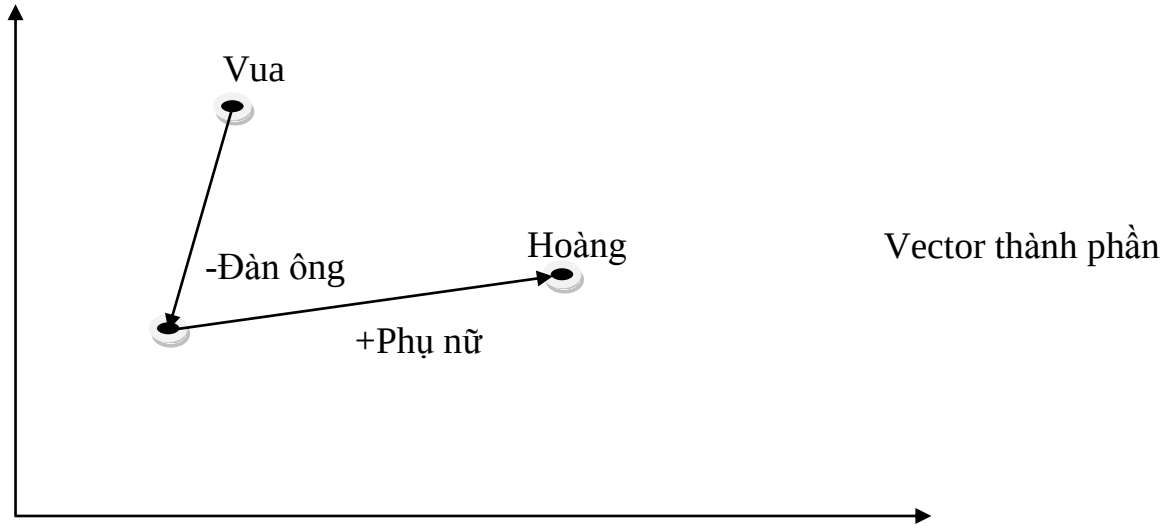
Hình 2.2: Mối quan hệ giữa số nhiều và số ít

Đây là sự hợp thành vector cũng cho phép ta trả lời câu hỏi "Vua – Đàn ông + Phụ nữ =?" và đi đến kết quả "Hoàng hậu"! Tất cả đều thực sự đáng chú ý khi bạn nghĩ rằng các kiến thức này chỉ đơn giản là xuất phát từ việc nhìn vào rất nhiều từ trong ngữ cảnh (ta sẽ thấy ngay) mà không có thông tin khác được cung cấp về ngữ nghĩa của nó.

Khá là ngạc nhiên để nhận thấy rằng sự giống nhau của các đại diện từ nằm ngoài các quy luật ngữ nghĩa đơn giản. Sử dụng kỹ thuật về giá trị bù từ nơi các phép toán đại số đơn giản được thực hiện trên các vector từ, điều đó đã được chỉ ra, ví dụ vector ("Vua") - vector ("Đàn ông") + vector ("Phụ nữ") cho kết quả trong một vector gần nhất với đại diện vector của từ “Hoàng hậu”.



Hình 2.3: Vector từ cho Vua, Đàn ông, Hoàng hậu và Phụ nữ

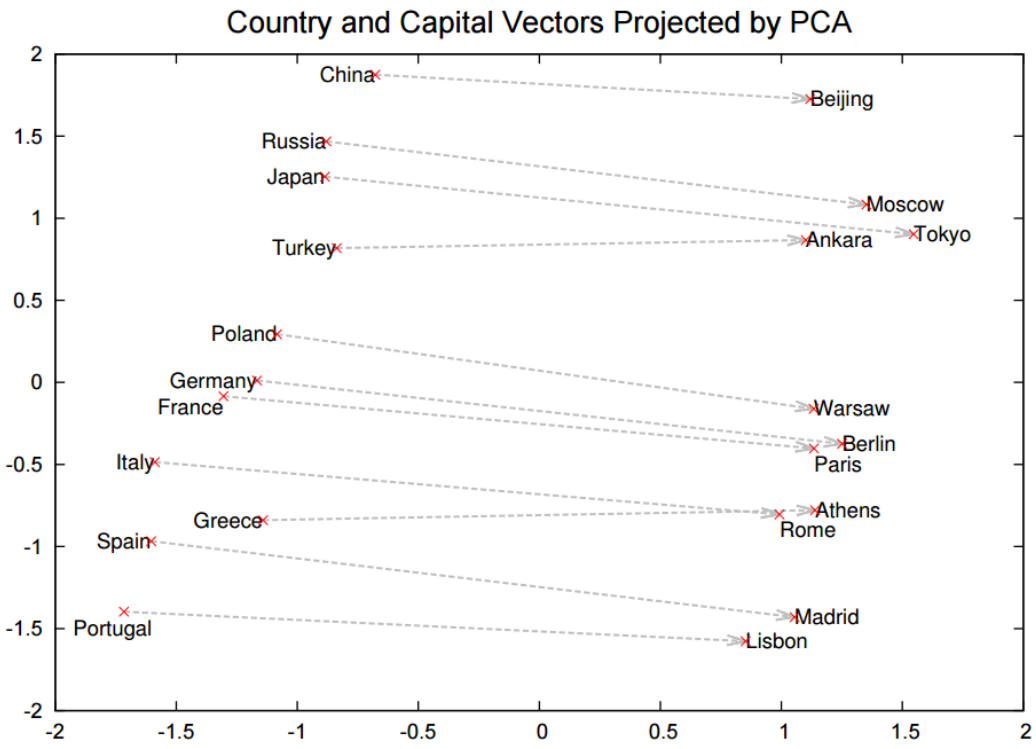


Hình 2.4: Kết quả sự cấu thành Vector Vua – Đàn ông + Phụ nữ = ?

Bảng 2.1: Ví dụ về các mối quan hệ giữ các cặp từ

Quan hệ	Ví dụ 1	Ví dụ 2	Ví dụ 3
France – Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
Big – bigger	Small: larger	Cold: colder	Quick: quicker
Miami – Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein – scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy – France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
Copper – Cu	Zinc: Zn	Gold: Au	Uranium: plutonium
Berlusconi – Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft – Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft – Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Dưới đây là mối quan hệ thủ đô-quốc gia (country-capital city) trông giống như 2 phép chiếu nhận diện hình ảnh 2 chiều:



Hình 2.5: Mối quan hệ thủ đô - quốc gia

Bảng 2.2: Ví dụ của các dạng câu hỏi "a là dành cho b như c là dành cho?"

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			

Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airline
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazor

Ta cũng có thể sử dụng thêm thành phần tương ứng của các thành phần vector để đặt câu hỏi chẳng hạn như 'Đức + các hãng hàng không" và bằng cách nhìn vào các dấu hiệu gần nhất với vector phức hợp đưa ra được câu trả lời ấn tượng:

Bảng 2.3: Trả lời cho câu hỏi dạng "a là dành cho b như c là dành cho?"

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
Koruna	Hanoi	Airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	Carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	Flag Carrier Lufthansa	Upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Vector từ với các mối quan hệ ngữ nghĩa như vậy có thể được sử dụng để cải thiện nhiều ứng dụng NLP hiện có, chẳng hạn như biên dịch bằng máy, hệ thống tìm kiếm thông tin và hệ thống câu hỏi/trả lời, và còn có thể cho phép các ứng dụng khác trong tương lai được phát minh.

Việc thử nghiệm mối quan hệ từ về ngữ nghĩa-cú pháp để hiểu về hàng loạt mối quan hệ như được thể hiện phía dưới. Sử dụng các Vector từ 640 chiều, mô hình skip-gram đạt được độ chính xác 55% về mặt ngữ nghĩa và 59% về mặt cú pháp.

2.3. Nghiên cứu các vector từ vựng

Mikolov và cộng sự [11] không phải là người đầu tiên sử dụng các đại diện vector liên tục của các từ, nhưng họ đã chỉ ra cách làm thế nào để giảm bớt sự phức tạp về mặt tính toán của việc nghiên cứu các đại diện như vậy - làm cho nó trở nên thực tế để nghiên cứu vector từ theo chiều cao trên một lượng cực lớn dữ liệu. Ví dụ, "Ta đã sử dụng một tập văn bản Tin tức Google để tạo các vector từ vựng. Tập văn bản này chứa khoảng 6 tỷ thẻ. Ta đã thu hẹp quy mô từ vựng đến 1 triệu từ quen thuộc nhất.."

Sự phức tạp trong các mô hình ngôn ngữ mạng neural (Truyền thẳng hay tái diễn) xuất phát từ lớp ẩn phi tuyến tính. Trong khi đây là những gì làm cho mạng neural trở nên rất hấp dẫn, vì vậy tôi quyết định tìm hiểu những mô hình đơn giản hơn, có thể không có khả năng đại diện cho các dữ liệu chính xác như các mạng neural, nhưng có thể được tạo trên nhiều dữ liệu hiệu quả hơn. Mikolov và cộng sự [11] đã đề xuất ra hai mô hình mới để sinh ra Word2Vec: Mô hình Continuous Bag-of-Words và mô hình Skip-gram.

2.4. Mô hình Continuous Bag-of-word/Mô hình túi từ liên tục (CBOW)

Mục tiêu huấn luyện của mô hình Continuous Bag-of-word là để dự đoán một từ khi biết các từ lân cận (ngữ cảnh) sử dụng mạng neural 3 tầng. Phần này tôi sẽ giới thiệu về ngữ cảnh của một từ và ngữ cảnh của một cụm từ.

2.4.1. Ngữ cảnh của một từ

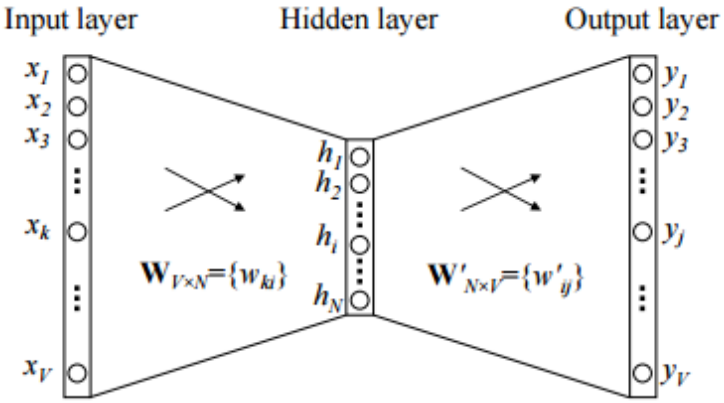
Ta bắt đầu từ phiên bản đơn giản nhất của mô hình CBOW được giới thiệu bởi Mikolov và cộng sự [11]. Ta giả định rằng chỉ có một từ được xem xét trong ngữ cảnh, có nghĩa là mô hình sẽ dự đoán một từ mục tiêu để xác định ngữ cảnh của từ, cái đó giống như mô hình Bigram.

Hình 2.6 sau đây biểu diễn mô hình mạng, sự định nghĩa ngữ cảnh đã được đơn giản hóa. Trong thiết lập của ta, quy mô từ vựng là V, và quy mô lớp ẩn là N. Các đơn vị trên lớp liền kề được kết nối đầy đủ Đầu vào là một vector được mã hóa one – hot, có nghĩa là cho một từ trong ngữ cảnh đầu vào được nhắc đến, chỉ có một trong số các đơn vị V, $\{x_1,\ldots,x_V\}$, sẽ là 1, và tất cả các đơn vị khác là 0.

Input layer: Lớp đầu vào

Hidden layer: Lớp ẩn

Output layer: Lớp đầu ra



Hình 2.6: Mô hình CBOW đơn giản với chỉ một từ trong ngữ cảnh

Các trọng số giữa lớp đầu vào và lớp đầu ra có thể được biểu diễn lại bằng một ma trận W kích thước V x N. Mỗi hàng của W là đại diện véc tơ N-chiều V_ω của từ liên kết của lớp đầu vào. Để xác định một ngữ cảnh (một từ), giả sử $x_k = 1$ và $x_{k'} = 0$ cho $k' \neq k$, theo đó:

$$h = \mathrm{W}^T x = \mathrm{W}^T_{(k,.)} := v^T_{\omega I'} \qquad (2.1)$$

trong đó chủ yếu là sao chép dòng thứ k của W tới h. $v_{\omega I}$ là đại diện vector của từ vựng đầu vào ω_I. Điều này ngụ ý rằng hàm liên kết (kích hoạt) của các đơn vị lớp ẩn là tuyến tính đơn giản (tức là, trực tiếp đi qua tổng trọng của đầu vào tới lớp tiếp theo).

Từ lớp ẩn tới lớp đầu ra, đó là một ma trận trọng số khác $\mathrm{W}'=\{\omega'_{ij}\}$, mà là một ma trận N x V. Sử dụng những trọng số này ta có thể tính toán một điểm u_j cho mỗi từ trong bộ từ vựng,

$$u_j = {v'_{\omega_j}}^T h \qquad (2.2)$$

với v'_{ω_j} là cột thứ j của ma trận W'. Sau đó, ta có thể sử dụng softmax, một mô hình phân lớp log-tuyến tính, để đạt được sự phân bố sau của các từ vựng, đây là sự phân phối đa thức.

$$p(\omega_j|\omega_I)=y_j=\frac{\exp(u_j)}{\sum_{j'=1}^{V}\exp(u_{j'})} \qquad (2.3)$$

trong đó y_j là đầu ra của đơn vị thứ j trong lớp đầu ra. Thay (2.1) và (2.2) vào (2.3), ta được:

$$p(\omega_j|\omega_I)=\frac{\exp(v'{\omega_j}^T v_{\omega_I})}{\sum_{j'=1}^{V}\exp(v'{\omega_{j'}}^T v_{\omega_I})} \qquad (2.4)$$

Lưu ý rằng v_ω và v'_ω là hai đại diện của từ ω. v_ω của dòng W, là đầu vào → ma trận trọng số ẩn, và v'_ω đến từ các cột của W' là ẩn → ma trận đầu ra. Trong phân tích tiếp theo, ta gọi v_ω là "vector đầu vào", và v'_ω như "vector đầu ra" của từ ω.

**** Cập nhật phương trình cho ẩn → trọng số***

Bây giờ ta suy ra phương trình cập nhật trọng số đối với mô hình này. Mặc dù việc tính toán hiện tại không thực tế (được giải thích phía dưới), ta đang suy luận để đạt được những hiểu biết về mô hình ban đầu này mà không có thủ thuật nào được áp dụng.

Mục tiêu huấn luyện (đối với một mẫu huấn luyện) là tối đa hóa (2.4), xác suất có điều kiện của việc quan sát từ đầu ra thực tế ω_0 (biểu thị chỉ số của nó trong lớp đầu ra như j*) được xác định nhóm các từ cùng ngữ cảnh đầu vào w_I chỉ quan tâm đến các trọng số. Ta đưa ra thuật toán tính xác suất có điều kiện và sử dụng nó để xác định hàm tổn thất.

$$\log p(\omega_O|\omega_I) = \log y_{j^*} \qquad (2.5)$$

$$= u_{j^*} - \log \sum_{j'=1}^{V} \exp(u_{j'}) := -E \qquad (2.6)$$

trong đó $E=-\log p(\omega_O|\omega_I)$ là hàm tổn thất, và j* là chỉ số của từ đầu ra thực tế. Lưu ý rằng hàm tổn thất có thể được hiểu như là một trường hợp đặc biệt của phép đo cross-entropy giữa hai phân phối xác suất.

Bây giờ ta lấy được các phương trình cập nhật của các trọng số giữa lớp ẩn và lớp đầu ra. Lấy đạo hàm của E đối với đầu vào u_j của đơn vị thứ j, ta được:

$$\frac{\partial E}{\partial u_j} = y_j - t_j := e_j \qquad (2.7)$$

Trong công thức (2.7) t_j=1(j=j*), tức là t_j sẽ là 1 trong khi các đơn vị thứ j là từ vựng đầu ra thực tế, nếu không t_j = 0. Lưu ý rằng đạo hàm này là lỗi dự đoán e_j của lớp đầu ra.

Tiếp theo ta lấy đạo hàm trên ω'_{ij} để có được độ chênh lệch trên các trọng số ẩn → các trọng số đầu ra:

$$\frac{\partial E}{\partial \omega'_{ij}} = \frac{\partial E}{\partial u_j} . \frac{\partial u_j}{\partial \omega'_{ij}} = e_j . h_i \qquad (2.8)$$

Vì vậy, sử dụng sự giảm độ chênh lệch ngẫu nhiên , ta được phương trình cập nhật trọng số cho ẩn → trọng số đầu ra:

$$\omega'^{(new)}_{ij} = \omega'^{(old)}_{ij} - \eta . e_j . h_i \qquad (2.9)$$

Hoặc:

$$v'^{(new)}_{\omega_j} = v'^{(old)}_{\omega_j} - \eta . e_j . h \qquad for\ j = 1,2,...,V \qquad (2.10)$$

Trong công thức trên η > 0 là tỷ lệ huấn luyện, e_j = y_j - t_j, và h_i là đơn vị thứ i trong lớp ẩn; v'_{ω_j} là vector đầu ra của ω_j. Lưu ý phương trình cập nhật này ngụ ý rằng ta phải đi qua tất cả các từ có thể trong lớp từ vựng, kiểm tra xác suất đầu ra y_j của nó, và so sánh y_j với xác suất đánh giá t_j (hoặc là 0 hoặc là 1). Nếu y_j> t_j (“đánh giá quá cao"), sau đó ta trừ một tỷ lệ h của vector ẩn (tức là: $\mathbf{v}_{\omega_I}$) từ $\mathbf{v'}_{\omega_j}$, rồi làm cho $\mathbf{v'}_{\omega_j}$ xa $\mathbf{v}_{\omega_I}$; nếu y_j < t_j (“đánh giá thấp"), ta thêm một số h cho $\mathbf{v'}_{\omega_O}$, sau đó làm cho $\mathbf{v'}_{\omega_O}$ gần $\mathbf{v}_{\omega I}$ hơn. Nếu y_j là rất gần với t_j rồi, căn cứ theo các phương trình cập nhật, rất ít thay đổi sẽ được thực hiện

đối với các trọng số. Lưu ý một lần nữa rằng $\mathbf{v}_\omega$ (vector đầu vào) và $\mathbf{v'}_\omega$ (vector đầu ra) là hai đại diện vector khác nhau của từ ω.

**** Cập nhật phương trình cho các trọng số đầu vào → trọng số ẩn***

Sau khi thu được các phương trình cập nhật cho W', bây giờ ta có thể chuyển sang W. Ta lấy đạo hàm của E ở đầu ra của các lớp ẩn, ta được:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} . \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} e_j . \omega'_{ij} := EH_i \qquad (2.11)$$

Trong công thức (2.11) h_i là đầu ra của đơn vị thứ i của lớp ẩn; u_j được định nghĩa trong (2.2), đầu vào thực của đơn vị thứ j trong lớp đầu ra; và $e_j = y_j - t_j$ là lỗi dự đoán của từ thứ j trong lớp đầu ra. EH, một vector N-chiều, là tổng của các vector đầu ra của tất cả các từ trong bộ từ vựng, được đánh trọng số bởi lỗi dự đoán của chúng.

Tiếp theo ta nên lấy đạo hàm của E trên W. Đầu tiên, nhớ lại rằng các lớp ẩn thực hiện một tính toán tuyến tính trên các giá trị từ lớp đầu vào. Mở rộng các ký hiệu vector trong (1.1), ta có được:

$$h_i = \sum_{k=1}^{V} x_k . \omega_{ki} \qquad (2.12)$$

Bây giờ ta lấy đạo hàm của E đối với mỗi phần tử của W, thì nhận được:

$$\frac{\partial E}{\partial \omega_{ki}} = \frac{\partial E}{\partial h_i} . \frac{\partial h_i}{\partial \omega_{ki}} = EH_i . x_k \qquad (2.13)$$

điều này tương đương với tích ten xơ (tensor) của x và EH, tức là:

$$\frac{\partial E}{\partial \mathrm{W}} = x \otimes EH = xEH^T \qquad (2.14)$$

từ đó ta có một ma trận kích thước V x N. Vì chỉ có một thành phần của x là khác 0, chỉ là một dòng của $\frac{\partial E}{\partial \mathrm{W}}$ khác 0, và giá trị của hàng đó là EH^T, và một vector N-chiều. Ta được phương trình cập nhật của W như sau:

$$v_{\omega_I}^{(new)} = v_{\omega_I}^{(old)} - \eta . EH^T \qquad (2.15)$$

Trong công thức (2.15) $\mathbf{v}_{\omega_I}$ là một hàng của W, "vector đầu vào" của nhóm từ cùng ngữ cảnh duy nhất, và là hàng duy nhất của W mà đạo hàm của nó khác 0. Tất cả các hàng khác của W sẽ vẫn không thay đổi sau sự lặp đi lặp lại này, bởi vì đạo hàm của chúng bằng 0.

Bằng trực giác, vì vector EH là tổng các vector đầu ra của tất cả các từ trong bộ từ vựng được đánh trọng số bởi lỗi dự đoán của chúng $e_j = y_j - t_j$, nên ta có thể hiểu (2.15) như thêm một phần của tất cả các vector đầu ra trong bộ từ vựng vào vector đầu vào của nhóm từ cùng ngữ cảnh. Nếu trong lớp đầu ra, xác suất của một từ ω_j là từ đầu ra được đánh giá quá cao ($y_j > t_j$), sau đó các vector đầu vào của nhóm từ cùng ngữ cảnh ω_I sẽ có xu hướng di chuyển ra xa vector đầu ra của ω_j; trái lại, nếu xác suất ω_j là từ đầu ra được đánh giá thấp ($y_j < t_j$), thì các vector đầu vào ω_I sẽ có xu hướng di chuyển gần hơn tới vector đầu ra của ω_j; nếu xác suất ω_j là dự đoán tương đối chính xác, thì nó sẽ có chút ảnh hưởng đến sự di chuyển của các vector đầu vào của ω_I. Sự di chuyển của vector đầu vào của ω_I được xác định bởi lỗi dự đoán của tất cả các vector trong vốn từ vựng; lỗi dự đoán càng lớn thì tác động càng lớn, một từ sẽ di chuyển trên vector đầu vào của nhóm từ cùng ngữ cảnh.

Vì ta cập nhật các thông số mô hình lặp đi lặp lại bằng việc bỏ qua cặp từ trong ngữ cảnh mục tiêu được tạo ra từ một tập huấn luyện, các kết quả trên các vector sẽ tích lũy. Ta có thể tưởng tượng rằng các vector đầu ra của một từ w bị "kéo" đi tới đi lui bởi các vector đầu vào của các từ đứng gần w cùng xảy ra, như thể có sợi dây vật lý giữa các vector của w và vector của các từ xung quanh nó. Tương tự như vậy, một vector đầu vào cũng có thể bị kéo bởi nhiều vector đầu ra. Việc giải thích này có thể nhắc nhở ta về lực hấp dẫn, hoặc sơ đồ đồ thị lực có hướng. Sau nhiều lần lặp lại, các vị trí tương đối của các vector đầu vào và đầu ra cuối cùng sẽ ổn định.

2.4.2. Ngữ cảnh của cụm từ

Hình 2.7 sau đây cho thấy mô hình CBOW với thiết lập ngữ cảnh của cụm từ. Khi tính toán đầu ra của lớp ẩn, thay vì trực tiếp sao chép vector đầu vào của nhóm từ cùng ngữ cảnh đầu vào, thì mô hình CBOW lấy trung bình các vector của các nhóm từ cùng ngữ cảnh đầu vào, và sử dụng các kết quả của ma trận trọng số đầu vào → ma trận trọng số ẩn và vector trung bình như đầu ra:

$$h = \frac{1}{C} \mathrm{W}^T (x_1 + x_2 + ... + x_C) \qquad (2.16)$$

$$= \frac{1}{C} (v_{\omega_1} + v_{\omega_2} + ... + v_{\omega_C})^T \qquad (2.17)$$

trong đó C là số các từ trong ngữ cảnh, ω_1; …; ω_C là các từ trong ngữ cảnh, và v_ω là vector đầu vào của một từ ω. Hàm tổn thất là:

$$E = -\log p(\omega_O | \omega_{I,1}, ..., \omega_{I,C} \qquad (2.18)$$

$$= -u_{j^*} + \log \sum_{j'=1}^{V} \exp(u_{j'}) \qquad (2.19)$$

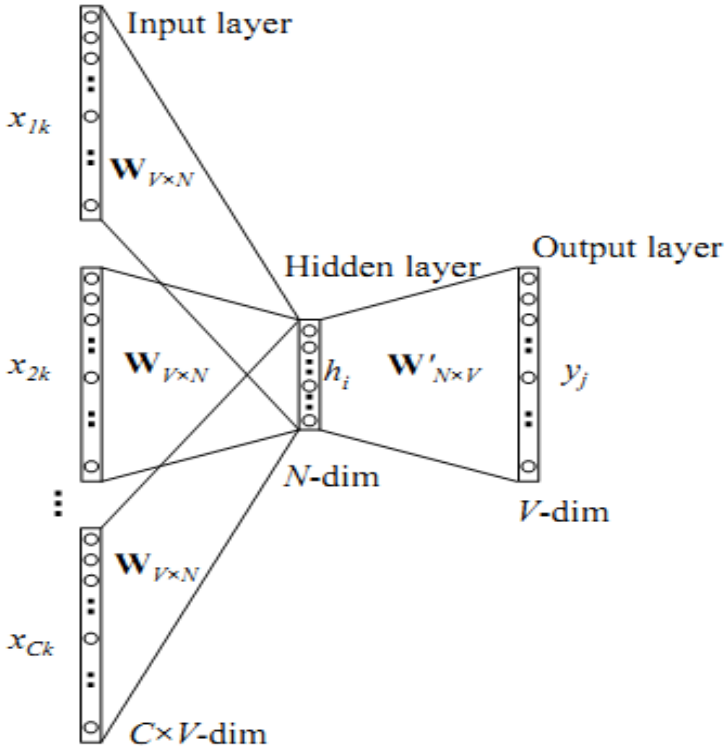
$$= -v'_{\omega o}{}^T . h + \log \sum_{j'=1}^{V} \exp(v'_{\omega_j}{}^T . h) \qquad (2.20)$$

giống như công thức (2.6), mục tiêu của mô hình one-word-context (từ một ngữ cảnh), ngoại trừ h là khác biệt, giống như định nghĩa trong công thức (2.17) thay vì công thức (2.1).

Input layer: Lớp đầu vào

Hidden layer: Lớp ẩn

Output layer: Lớp đầu ra



Hình 2.7: Mô hình túi từ liên tục (CBOW)

Phương trình cập nhật cho các trọng số ẩn → trọng số đầu ra là giống nhau đối với mô hình một từ trong ngữ cảnh (2.10). Tôi xin được chép lại dưới đây:

$$v'^{(new)}_{\omega_j} = v'^{(old)}_{\omega_j} - \eta.e_j.h \quad j=1,2,...,V \qquad (2.21)$$

Lưu ý rằng ta cần phải áp dụng điều này đối với mọi phần tử của ma trận trọng số ẩn → ma trận trọng số đầu ra cho mỗi ví dụ huấn luyện.

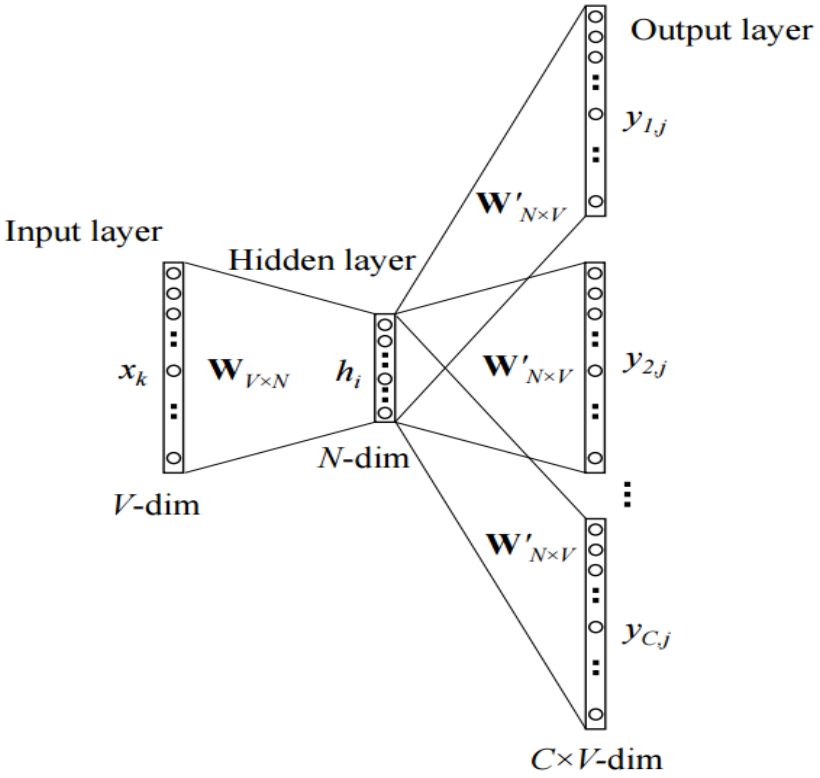
Các phương trình cập nhật cho trọng số đầu vào → trọng số ẩn tương tự công thức (2.15), ngoại trừ bây giờ ta cần phải áp dụng phương trình sau đây cho mỗi từ $\omega_{I,c}$ trong ngữ cảnh:

$$v^{(new)}_{\omega_{I,c}} = v^{(old)}_{\omega_{I,c}} - \frac{1}{C}.\eta.EH^T \quad c=1,2,...,C \qquad (2.22)$$

Trong công thức (2.22) $\mathbf{v}_{\omega_{I,c}}$ là véc tơ đầu vào của từ vựng thứ c trong ngữ cảnh đầu vào; η là tỷ lệ huấn luyện tích cực và $EH=\frac{\partial E}{\partial h_i}$ được nhắc đến trong công thức (2.11). Sự hiểu trực giác của phương trình cập nhật tương tự như công thức (2.15).

2.5. Mô hình Skip-gram

Mô hình Skip-gram được đưa ra bởi Mikolov và các cộng sự [10,11]. Mô hình này trái ngược lại với mô hình CBOW. Các từ mục tiêu bây giờ lại ở lớp đầu vào và các từ cùng ngữ cảnh lại ở lớp đầu ra.



Hình 2.8: Mô hình Skip-gram

Mục tiêu huấn luyện của mô hình Skip-gram là để tìm ra đại diện từ vựng hữu ích để dự đoán các từ xung quanh trong một câu hay một tài liệu. Chính thức hơn, đưa ra một chuỗi các từ huấn luyện $\omega_1, \omega_2, \omega_3, \ldots, \omega_T$, mục tiêu của mô hình Skip-gram là tối đa hóa xác suất log trung bình. Ta có công thức:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(\omega_{t+j}|\omega_t) \qquad (2.23)$$

trong đó c là quy mô của ngữ cảnh huấn luyện (mà có thể là một hàm số của từ trung tâm ω_T). Khi sử dụng khoảng thời gian huấn luyện, kết quả c lớn hơn trong nhiều ví dụ huấn luyện và do đó có thể dẫn đến một độ chính xác cao hơn. Việc xây dựng Skip-gam cơ bản xác định p (ω_{t+j} | ω_t) bằng cách sử dụng hàm softmax:

$$p(\omega_O|\omega_I = \frac{\exp(v'_{\omega_O}{}^T v_{\omega_I})}{\sum_{\omega=1}^{W}\exp(v'_{\omega}{}^T v_{\omega_I})} \qquad (2.24)$$

Trong công thức (2.24) v_ω và v'_ω là các đại diện vector "đầu vào" và "đầu ra" của ω và W là số từ trong bộ từ vựng. Công thức này không thực tế bởi vì giá trị của phép tính ∇log p ($\omega_O|\omega_I$) tỷ lệ thuận với W, mà giá trị này thì thường rất lớn (10^5- 10^7).

2.5.1. Hierarchical Softmax (Softmax phân cấp)

Một phép tính xấp xỉ hiệu quả của toàn bộ softmax là Softmax phân cấp. Trong ngữ cảnh của các mô hình ngôn ngữ mạng neural, nó được giới thiệu lần đầu tiên bởi Morin và Bengio [8]. Những ưu điểm chính là thay vì đánh giá các nút đầu ra W trong mạng neural để nhận được phân bố xác suất, nó là cần thiết cho việc đánh giá về các nút $\log_2(W)$ duy nhất.

Đối với mỗi nút, Softmax phân cấp sử dụng một cây đại diện nhị phân của lớp đầu ra với các từ W như lá của nó, đối với mỗi nút, rõ ràng các đại diện xác suất tương đối của các nút con của nó. Những điều này định nghĩa một bước đi ngẫu nhiên được cho là xác suất đối với các từ.

Chính xác hơn, mỗi từ ω có thể đạt được bằng một đường từ gốc của cây. Gọi n(ω, j) là nút thứ j trên con đường từ gốc đến ω và gọi L(ω) là độ dài của đường đi đó, thì n(ω,1) = gốc và n(ω,L(ω)) = ω. Hơn nữa, đối với bất kỳ nút bên trong n nào, gọi ch(n) là tập con tùy ý và gọi x là 1 nếu x là đúng và sai là -1. Vậy Softmax phân cấp xác định $p(\omega_o|\omega_I)$ như sau:

$$p(\omega|\omega_I)=\prod_{j=1}^{L(\omega)-1}\sigma(\ n(\omega,j+1)=ch(n(\omega,j))\ .v'_{n(\omega,j)}{}^{T}v_{\omega_I}) \qquad (2.25)$$

Trong công thức (2.25) $\sigma(x)=1/(1+\exp(-x))$ có thể được xác định rằng $\sum_{\omega=1}^{W} p(\omega|\omega_I)=1$. Điều này ngầm chỉ ra rằng trị giá của phép tính $\log p(\omega_O|\omega_I)\ and\ \nabla \log p(\omega_O|\omega_I)$ là tỷ lệ thuận với $L(\omega_o)$, trị giá trung bình không lớn hơn logW. Cũng không giống như công thức softmax chuẩn của Skip-gram mà gán hai đại diện v_ω và v'_ω đối với mỗi từ ω, công thức Softmax phân cấp có một đại diện v_ω đối với mỗi từ ω và một đại diện v'_n đối với mỗi nút trong n của cây nhị phân.

Cấu trúc của cây được sử dụng bởi softmax phân cấp có tác dụng đáng kể về hiệu suất. Mnih và Hinton đã khám phá một số phương pháp để xây dựng các cấu trúc cây và các hiệu ứng trên cả thời gian huấn luyện và tính chính xác của mô hình kết quả [5]. Trong công trình của họ sử dụng một cây Huffman nhị phân, như nó gán mã ngắn đối với các từ thường gặp mà tạo kết quả nhanh. Nó đã được quan sát trước khi nhóm các từ với nhau bằng tần suất của chúng hoạt động tốt như một kỹ thuật tăng tốc đơn giản cho mạng neural dựa trên các mô hình ngôn ngữ [11,12].

2.5.2. Negative Sampling (Mẫu phủ định)

Một thay thế cho softmax phân cấp là Noise Contrastive Estimation (NCE - Ước tính tương phản nhiễu), được Gutmann và Hyvarinen giới thiệu [9] và Mnih và Teh đã áp dụng cho mô hình ngôn ngữ [6]. NCE thừa nhận rằng một mô hình tốt nên có khả năng phân biệt dữ liệu nhiễu bằng các phương tiện hồi quy logistic. Điều này cũng tương tự như việc mất đi điểm mấu chốt mà Collobert và Weston đã sử dụng [14] họ là những người huấn luyện các mô hình bằng cách xếp hạng các dữ liệu nhiễu.

Trong khi NCE có thể được hiển thị để tối đa hóa xác suất log của softmax, thì mô hình Skip-gram lại chỉ quan tâm đến việc nghiên cứu đại diện vector chất lượng cao, vì vậy ta được tự do để đơn giản hóa NCE miễn là các đại diện vector giữ được chất lượng của chúng. Ta xác định lấy mẫu phủ định (NEG) là mục tiêu:

$$\log\sigma(v'_{\omega_O}{}^{T} v_{\omega_I}) + \sum_{i=1}^{k} \mathrm{E}_{\omega_i \sim P_n(\omega)} \left[\log\sigma(-v'_{\omega_i}{}^{T} v_{\omega_I}) \right] \qquad (2.26)$$

2.5.3. Subsampling of Frequent Words (Lựa chọn mẫu phụ của các từ thường gặp).

Trong một tập văn lớn, các từ thường thấy nhất có thể dễ gặp hàng trăm triệu lần (ví dụ, “in”, “the”, và “a”). Những từ như vậy thường cung cấp giá trị thông tin ít hơn những từ hiếm gặp. Ví dụ, trong khi những lợi ích mô hình Skip-gram từ việc quan sát các sự xuất hiện đồng thời của "France" và "Paris", nó giúp ích ít nhiều từ việc quan sát sự đồng xuất hiện thường xuyên của "France" và "the", như hầu hết các từ cùng xuất hiện thường xuyên trong một câu với "the". Ý tưởng này cũng có thể được áp dụng theo hướng ngược lại; các đại diện vector của các từ thường gặp không làm thay đổi đáng kể sau khi thực hiện trên vài triệu ví dụ.

Để tránh sự mất cân bằng giữa các từ hiếm và thường gặp, ta đã sử dụng một phương pháp tiếp cận mẫu phụ đơn giản: mỗi từ ω_i trong tập huấn luyện được loại bỏ với xác suất tính theo công thức:

$$P(\omega_i) = 1 - \sqrt{\frac{t}{f(\omega_i)}} \qquad (2.27)$$

Bảng 2.4: Độ chính xác của nhiều mô hình Skip-gram 300-chiều

Phương pháp	Thời gian (phút)	Cú pháp (%)	Ngữ nghĩa (%)	Tổng độ chính xác (%)
NEG-5	38	63	54	59
NEG-15	97	63	58	61
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
Nhưng kết quả sau sử dụng 10^{-5} mẫu phụ				
NEG-5	14	61	58	60
NEG-15	36	61	61	61

HS-Huffman	21	52	59	55

trong đó f (ω_i) là tần số của từ ω_i và t là một ngưỡng được chọn, thường khoảng 10^{-5}. Ta đã lựa chọn công thức mẫu phụ này vì nó cho thấy các từ mẫu phụ có tần số lớn hơn t trong khi vẫn giữ thứ hạng của các tần số. Mặc dù công thức mẫu phụ này đã được lựa chọn một cách kín đáo nhưng ta đã ứng dụng rất ổn trong thực tế. Nó làm tăng tốc việc nghiên cứu và thậm chí cải thiện đáng kể độ chính xác của các vector đã được nghiên cứu của những từ hiếm gặp.

CHƯƠNG 3: ỨNG DỤNG WORD2VEC VÀO PHÂN LOẠI GIỚI TÍNH NGƯỜI DÙNG MẠNG XÃ HỘI

3.1. Mở đầu

Ngày nay, với sự phát triển không ngừng của khoa học kỹ thuật, công nghệ thông tin trên thế giới nói chung và ở Việt Nam nói riêng có những bước tiến vượt bậc. Song hành với nó là sự phát triển của cơ sở hạ tầng, trang thiết bị công nghệ thông tin. Theo báo cáo tổng kết công tác năm 2015 và phương hướng, nhiệm vụ năm 2016 của Cục Viễn thông (Bộ Thông tin và Truyền thông), tính tới thời điểm cuối năm 2015, Việt Nam có Tỷ lệ người dùng Internet tại Việt Nam đã đạt 52% dân số. Internet băng rộng di động có 36,28 triệu thuê bao, với tỷ lệ 40,1 thuê bao/100 dân.

Bên cạnh đó, theo thống kê của "wearesocial.net", tháng 1-2015, người Việt Nam đang đứng thứ 4 trên thế giới về thời gian sử dụng Internet với 5,2 giờ mỗi ngày, chỉ sau Philippines đứng đầu là 6 giờ, tiếp đó là Thái lan với 5,5 giờ, và Brazin là 5,4 giờ/ngày. Hơn nữa, người Việt Nam sử dụng Internet cũng đứng thứ 9 về số thời gian trung bình dành cho mạng xã hội là 3,1giờ mỗi ngày; đứng thứ 22 trên thế giới tính theo dân số về số người sử dụng mạng xã hội là 31%, trong đó facebook là mạng xã hội được sử dụng thông dụng nhất.

Chính vì sự phát triển không ngừng của công nghệ thông tin và mức độ phổ biến của mạng xã hội ngày nay mà thông tin trên mạng xã hội vô cùng phong phú và liên tục. Người dùng mạng xã hội ở Việt Nam hiện nay thường có thói quen cung cấp, trao đổi các thông tin một cách liên tục và thường xuyên. Hầu hết các thông tin được trao đổi hiện nay nằm dưới dạng tài liệu văn bản. Các thông tin đó có thể là các bài báo, các tài liệu kinh doanh, các thông tin kinh tế, các bài nghiên cứu khoa học, các thông tin cá nhân khác,... Từ thực tế đó đã xuất hiện các nhu cầu phân tích thông tin để phân loại các thông tin đó cho các mục đích khác nhau như học tập, nghiên cứu, kinh doanh, ...

Với thực tế đó, vấn đề đặt ra là làm thế nào ta có thể khai thác được những thông tin hữu ích từ các nguồn dữ liệu phong phú của mạng xã hội. Các nguồn dữ liệu này phải được xử lý như thế nào để người dùng có thể có những công cụ tự động hoá trợ giúp trong việc phát hiện tri thức và khai thác thông tin. Rõ ràng, ta phải hiểu rõ bản chất của dữ liệu văn bản, hiểu rõ các đặc trưng của các dữ liệu loại này để có thể có được những phương pháp luận cần thiết.

Việc khai thác thông tin từ các nguồn dữ liệu văn bản trên mạng xã hội ở Việt Nam chắc chắn phải dựa vào những kết quả nghiên cứu về văn bản nói chung, về dữ liệu văn bản và các kỹ thuật xử lý đã được phát triển trên thế giới. Tuy nhiên, những văn bản tiếng Việt lại có những đặc trưng riêng của nó. Ta có thể nhận thấy được ngay sự khác biệt về mặt kí pháp, cú pháp và ngữ pháp tiếng Việt trong các văn bản so với các ngôn ngữ phổ biến trên thế giới như tiếng Anh, tiếng Pháp. Vậy thì những đặc trưng này ảnh hưởng thế nào đến các kỹ thuật khai phá dữ liệu văn bản, ta cần phải có những kỹ thuật mới nào để có thể tận dụng được những ưu thế của tiếng Việt cũng như giải quyết được những phức tạp trong tiếng Việt.

Hiện nay, đã xuất hiện một số phương pháp và kỹ thuật khai phá dữ liệu văn bản tiếng Việt tuy nhiên với mỗi phương pháp đều có ưu, nhược điểm khác nhau. Trong mỗi lĩnh vực khác nhau thì các phương pháp cũng cho kết quả phân tích, phân loại là khác nhau. Và để cải thiện kết quả của việc khai phá dữ liệu văn bản tiếng Việt tôi đã đề xuất sử dụng Word2Vec để đưa vào quá trình phân tích và phân loại cho văn bản.

Đối với các kỹ thuật không sử dụng Word2Vec, khi phân loại văn bản người ta sẽ trích chọn các đặc trưng tương ứng với các nhãn được gán. Sau đó các đặc trưng này sẽ được lưu vào bộ từ vựng cùng với các trọng số của nó. Tuy nhiên bộ từ vựng này sẽ có thể xảy ra tình trạng thưa dữ liệu và vấn đề kích thước của mô hình ngôn ngữ là rất lớn nếu tập văn bản có kích thước lớn. Còn đối với kỹ thuật có sử dụng thêm Word2Vec, các từ sẽ được biểu diễn bởi các vector. Các từ có ý nghĩa gần giống nhau thì có kích thước vector gần bằng nhau. Chính vì vậy mà Word2Vec có thể tự học được đối với các từ cùng ngữ cảnh.

Để dễ hình dung quá trình khai phá dữ liệu văn bản tiếng Việt, tôi đưa ra bài toán "Phân loại giới tính người dùng mạng xã hội dựa vào tin nhắn văn bản và Word2Vec".

3.2. Giải pháp cho bài toán phân loại giới tính người dùng mạng xã hội

Đối với bài toán phân loại giới tính người dùng mạng xã hội tôi sử dụng phương pháp học máy để xử lý. Trong học máy có hai vấn đề cần được giải quyết, đó là:

**** Lựa chọn bộ phân lớp:***

Trong quá trình thực nghiệm tôi đã áp dụng bài toán này trên cả hai bộ phân lớp phổ biến và được đánh giá hiệu quả hiện nay là bộ phân lớp SVM và Logistic. Kết quả thu được cho thấy bộ phân lớp Logistic cho kết quả thực nghiệm tốt hơn. Chính vì vậy mà tôi đã lựa chọn bộ phân lớp Logistic để sử dụng cho bài toán phân loại giới tính người dùng mạng xã hội. Tuy nhiên tôi không đi sâu vào tìm hiểu cách làm việc của bộ phân lớp này mà chỉ sử dụng công cụ GraphLab Create[5].

Bộ phân lớp Logistic được sử dụng cho mô hình hồi qui Logistic. Bằng cách trích chọn một tập các trọng số của các đặc trưng từ đầu vào, lấy các bản ghi, và kết hợp chúng tuyến tính (có nghĩa là mỗi đặc trưng được nhân với một trọng số và sau đó cộng lại). Mục tiêu hồi qui Logistic là nghiên cứu mối tương quan giữa một (hay nhiều) yếu tố nguy cơ (risk factor) và đối tượng phân tích (outcome). Chẳng hạn như đối với nghiên cứu mối tương quan giữa thói quen hút thuốc lá và nguy cơ mắc ung thư phổi thì yếu tố nguy cơ ở đây là thói quen hút thuốc lá và đối tượng phân tích ở đây là nguy cơ mắc ung thư phổi. Trong hồi qui logistic thì các đối tượng nghiên cứu thường được thể hiện qua các biến số nhị phân (binary) như xảy ra/không xảy ra; chết/sống; có/không; nam/nữ;… còn các yếu tố nguy cơ có thể được thể hiện qua các biến nhị phân (giới tính) hay các biến thứ bậc (thu nhập: Cao, trung bình, thấp). Vấn đề đặt ra cho nghiên cứu dạng này là làm sao để ước tính độ tương quan của các yếu tố nguy cơ và đối tượng phân tích.

**** Trích chọn đặc trưng:***

Khi đã có được một bộ phân lớp tốt thì việc phân loại hiện giờ sẽ phụ thuộc rất nhiều vào các đặc trưng đưa vào để phân loại. Đặc trưng càng chính xác thì việc phân loại càng nhận được kết quả tốt. Chính vì vậy mà việc trích chọn đặc trưng vô cùng quan trọng. Đối với quá trình thực nghiệm trong luận văn này tôi chủ yếu tìm hiểu để xây dựng và trích chọn được các đặc trưng tốt nhằm cải thiện kết quả phân loại. Do đó mà bước lấy dữ liệu thô, sau đó tiền xử lý và xây dựng các bộ dữ liệu sẽ quyết định rất nhiều đối với việc trích chọn đặc trưng. Dữ liệu càng mịn, càng ít nhiễu thì đặc trưng càng chính xác. Đối với bài

[5] https://turi.com/products/create/

toán phân lớp văn bản ban đầu sẽ chọn đặc trưng theo mô hình n-gram với n=1,2,3. Sau đó các đặc trưng này sẽ được sử dụng cho bộ phân lớp.

Khi bộ phân lớp thực hiện việc phân lớp nó sẽ sử dụng các đặc trưng lấy được từ tập dữ liệu kiểm thử rồi tìm liên kết đến các đặc trưng được trích chọn từ tập dữ liệu huấn luyện theo mô hình n-gram. Tuy nhiên khi sử dụng kỹ thuật này hay xảy ra tình trạng thưa dữ liệu, phân bố không đồng đều. Bên cạnh đó, khi kích thước tập văn bản huấn luyện lớn, số lượng các cụm n-gram và kích thước của mô hình ngôn ngữ cũng rất lớn. Chính vì vậy tôi sử dụng thêm Word2Vec để đưa thêm được ngữ cảnh từ vào cho các đặc trưng.

Giả sử khi các đặc trưng của tập dữ liệu kiểm thử không tìm thấy bất kỳ một liên kết nào với các đặc trưng đã được trích chọn trong tập dữ liệu huấn luyện. Khi đó bộ phân lớp sẽ tìm kiếm trong Word2Vec các từ có nghĩa gần với các từ của các đặc trưng trong tập dữ liệu kiểm thử, sau khi tìm được các từ gần nghĩa hoặc giống nhau nó sẽ coi đây là các đặc trưng của bộ dữ liệu kiểm thử và đưa vào tìm các liên kết với các đặc trưng được trích chọn của tập huấn luyện. Chính vì có khả năng tự học được các từ có nghĩa tương đồng như vậy mà khi sử dụng thêm Word2Vec thì kết quả phân lớp của chúng ta sẽ tăng lên.

3.2.1. Phân loại theo mô hình n-gram

Mô hình ngôn ngữ là một phân bố xác suất trên các tập văn bản. Nói đơn giản, mô hình ngôn ngữ có thể cho biết xác suất một câu (hoặc cụm từ) thuộc một ngôn ngữ là bao nhiêu [2].

Ví dụ: khi áp dụng mô hình ngôn ngữ cho tiếng Việt:

P[“hôm qua là thứ năm”] = 0.001

P[“năm thứ hôm là qua”] = 0

Mô hình ngôn ngữ được áp dụng trong rất nhiều lĩnh vực của xử lý ngôn ngữ tự nhiên như: kiểm lỗi chính tả, dịch máy hay phân đoạn từ... Chính vì vậy, nghiên cứu mô hình ngôn ngữ chính là tiền đề để nghiên cứu các vấn đề, bài viết tiếp theo trong xử lý ngôn ngữ tự nhiên.

Mô hình ngôn ngữ có nhiều hướng tiếp cận, nhưng chủ yếu được xây dựng theo mô hình n-gram.

Khái niệm về n-gram: là tần suất xuất hiện của n kí tự (hoặc từ) liên tiếp nhau có trong dữ liệu của kho ngữ liệu (*corpus*[6]).

Với n = 1 và tính trên kí tự, ta có thông tin về tần suất xuất hiện nhiều nhất của các chữ cái. Điều này ứng dụng để bố trí các phím trên bàn phím máy tính: các phím hay xuất hiện nhất sẽ ở những vị trí dễ sử dụng nhất.

Với n = 2, ta có khái niệm bigram. Ví dụ với các chữ cái tiếng Anh, 'th','he','in','an','er' là các cặp kí tự hay xuất hiện nhất. Ngoài ra, ta có thể biết thêm rằng sau kí tự 'q' thì phần lớn đều là kí tự 'u'.

Với n = 3, ta có trigram. Nhưng vì n càng lớn thì số trường hợp càng lớn nên thường người ta chỉ sử dụng với n = 1,2 hoặc đôi lúc là 3. Ví dụ với các kí tự tiếng Anh, tiếng Anh sử dụng 26 kí tự, vậy với n = 1 thì số trường hợp là 26, n = 2 thì số trường hợp là 26^2 = 676 trường hợp, n = 3 có 17576 trường hợp.

Bigram được sử dụng nhiều trong việc phân tích hình thái (từ, cụm từ, từ loại) cho các ngôn ngữ khó phân tích như tiếng Việt, tiếng Nhật, tiếng Trung, … Dựa vào tần suất xuất hiện cạnh nhau của các từ, người ta sẽ tính cách chia 1 câu thành các từ sao cho tổng bigram là cao nhất có thể. Với thuật giải phân tích hình thái dựa vào trọng số nhỏ nhất, người ta sử dụng n = 1 để xác định tần suất xuất hiện của các từ và tính trọng số.

Do đó, để đảm bảo tính thống kê chính xác đòi hỏi các corpus phải lớn và có tính đại diện cao.

**** Áp dụng mô hình n-gram cho bài toán phân loại giới tính người dùng mạng xã hội ta thực hiện như sau:***

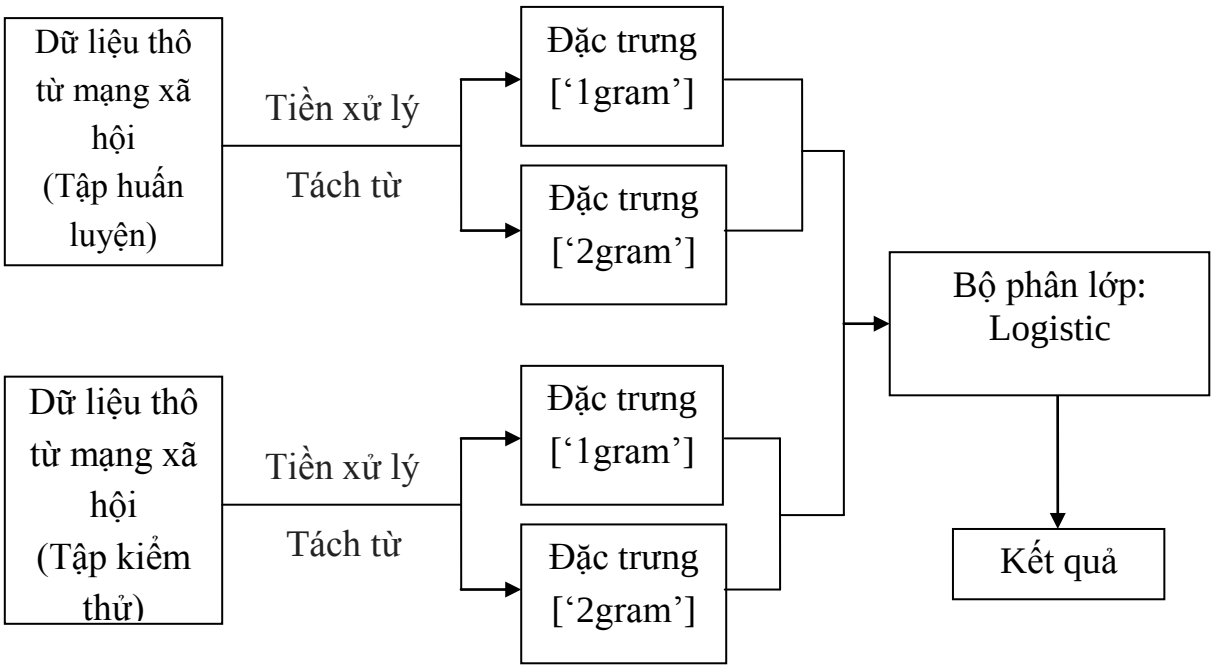
Ban đầu tôi lấy dữ liệu trên tường của từng người dùng mạng xã hội lưu thành một file và gán nhãn cho dữ liệu này theo thông tin trên tài khoản của họ là nam hay nữ. Sau đó tiền xử lý dữ liệu rồi sử dụng công cụ vn.vitk-master để thực hiện tách từ và thu được các dữ liệu đã chuẩn hóa. Từ dữ liệu đã chuẩn hóa tôi trích chọn đặc trưng 1-gram và 2-gram cùng giá trị nhãn đã được gán để sử dụng làm tập dữ liệu huấn luyện. Nghĩa là, các đặc trưng của tập dữ liệu huấn luyện được phân thành hai lớp giới tính nam và giới tính nữ.

[6] http://viet.jnlp.org/tai-nguyen-ngon-ngu-tieng-viet/khai-yeu-ve-corpus

Trong quá trình trích chọn đặc trưng (1-gram và 2-gram) đối với từng người dùng tôi sẽ tìm xem các từ người dùng hay sử dụng (đã loại bỏ các từ vô nghĩa - stopword) và đưa vào làm đặc trưng cho lớp tương ứng đã được gán nhãn của người dùng đó. Ví dụ với một người dùng tôi thấy hay sử dụng cụm từ "đi đá_bóng" và người này đã được gán nhãn là Nam thì tôi sẽ đưa cụm từ này vào làm đặc trưng cho lớp giới tính Nam của tập huấn luyện. Với một người dùng tôi thấy hay sử dụng cụm từ "đi mua_sắm" và người này đã được gán nhãn là Nữ thì tôi sẽ đưa cụm từ "đi mua_sắm" vào làm đặc trưng cho lớp giới tính Nữ của tập huấn luyện. Tương tự như vậy đối với các người dùng khác thì tôi sẽ xây dựng được một tập dữ liệu huấn luyện chứa các đặc trưng đã được phân lớp.

Để phân loại giới tính của một người dùng tôi sẽ đưa lần lượt từng đặc trưng đã trích chọn trên tường của người này (dữ liệu kiểm thử) vào bộ phân lớp Logistic để so sánh với các đặc trưng trong tập dữ liệu huấn. Bộ phân lớp này sẽ tìm kiếm một liên kết giữa đặc trưng của dữ liệu kiểm thử với các đặc trưng trong tập huấn luyện. Sau đó sẽ kiểm tra xem đặc trưng trong tập huấn luyện được tìm thấy nằm ở lớp nào sẽ có được kết quả. Sau đó lại tiếp tục với các đặc trưng còn lại của dữ liệu kiểm thử. Cuối cùng giới tính của người dùng này sẽ được xác định dựa vào những kết quả thu được. Kết quả đặc trưng nằm ở lớp nào chiếm đa số sẽ được sử dụng làm kết quả cuối cùng.

Để dễ hình dung quy trình tôi có sơ đồ như sau:



Hình 3.1: Phân loại theo mô hình n-gram

3.2.2. Phân loại khi sử dụng thêm Word2Vec

Mô hình n-gram ở trên hiện đang được sử dụng phổ biến trong xử lý ngôn ngữ tự nhiên. Tuy nhiên nó lại gặp phải một số khó khăn.

Một là, dữ liệu phân bố không đều. Khi sử dụng mô hình n-gram theo công thức "xác suất thô", sự phân bố không đều trong tập văn bản huấn luyện có thể dẫn đến các ước lượng không chính xác. Khi các n-gram phân bố thưa, nhiều cụm n-gram không xuất hiện hoặc chỉ có số lần xuất hiện nhỏ, việc ước lượng các câu có chứa các cụm n-gram này sẽ có kết quả không cao. Với V là kích thước bộ từ vựng, ta sẽ có V cụm n-gram có thể sinh từ bộ từ vựng. Tuy nhiên, thực tế thì số cụm n-gram có nghĩa và thường gặp chỉ chiếm rất ít.

Ví dụ: Tiếng Việt có khoảng hơn 5000 âm tiết khác nhau, ta có tổng số cụm 3-gram có thể có là: 5.000^3 = 125.000.000.000. Tuy nhiên, số cụm 3-gram thống kê được chỉ xấp xỉ 1.500.000. Như vậy sẽ có rất nhiều cụm 3-gram không xuất hiện hoặc chỉ xuất hiện rất ít.

Khi tính toán xác suất của một câu, có rất nhiều trường hợp sẽ gặp cụm n-gram chưa xuất hiện trong dữ liệu huấn luyện bao giờ. Điều này làm xác suất của cả câu bằng 0, trong khi câu đó có thể là một câu hoàn toàn đúng về mặt ngữ pháp và ngữ nghĩa.

Hai là, kích thước bộ nhớ của mô hình ngôn ngữ lớn. Khi kích thước tập văn bản huấn luyện lớn, số lượng các cụm n-gram và kích thước của mô hình ngôn ngữ cũng rất lớn. Nó không những gây khó khăn trong việc lưu trữ mà còn làm tốc độ xử lý của mô hình ngôn ngữ giảm xuống do bộ nhớ của máy tính là hạn chế. Để xây dựng mô hình ngôn ngữ hiệu quả, chúng ta phải giảm kích thước của mô hình ngôn ngữ mà vẫn đảm bảo độ chính xác.

Ta xét câu sau: "Tôi thích đá bóng". Giả sử cụm từ "thích đá bóng" đã được lưu trong các đặc trưng được trích chọn. Khi đó nếu sử dụng mô hình n-gram phân loại thì sẽ cho xác suất cao là Nam giới. Tuy nhiên với câu sau: "Tôi đam mê đá bóng" nhưng trong bộ đặc trưng được trích chọn lại không có cụm từ "đam mê đá bóng". Khi đó nếu sử dụng mô hình n-gram bình thường để phân loại thì sẽ không phân loại được hoặc phân loại không chính xác. Chính vì vậy mà tôi đề xuất sử dụng thêm Word2Vec làm đặc trưng cho trường hợp này. Trong Word2Vec các từ gần giống nhau thì giá trị của các vector từ đó là gần như nhau. Ví dụ như khoảng cách giữa các vector từ "thích" và vector từ "đam

mê" là gần bằng nhau. Chính vì vậy khi ta không sử dụng Word2Vec thì việc biểu diễn mô hình ngôn ngữ của chúng ta đòi hỏi kích thước rất lớn.

**** Áp dụng phân loại khi đưa thêm Word2Vec làm đặc trưng cho mô hình n-gram ta thực hiện như sau:***

Để có thể cải thiện những hạn chế nêu trên ta sẽ sử dụng Word2Vec biểu diễn cho các từ về dạng vector. Đầu tiên tôi lấy nội dung các bài báo trên các trang web như; 24h.com.vn, vnexpress.net, eva.vn, dantri.vn,... Sau đó tiến hành tiền xử lý rồi sử dụng công cụ vn.vitk-master để thực hiện tách từ và thu được dữ liệu đã chuẩn hóa. Tiếp theo tôi sử dụng công cụ Gensim[7] cho bộ dữ liệu đã chuẩn hóa để sinh ra một file Word2Vec dùng làm đặc trưng vector. Trong đặc trưng vector này các từ có nghĩa gần nhau sẽ được biểu diễn với giá trị gần bằng nhau. Ví dụ: từ xe_máy có giá trị biểu diễn là 0.7628448512386902 sẽ có các từ sau được biểu diễn có giá trị gần như nhau:

Bảng 3.1: Giá trị biểu diễn các từ trong Word2Vec

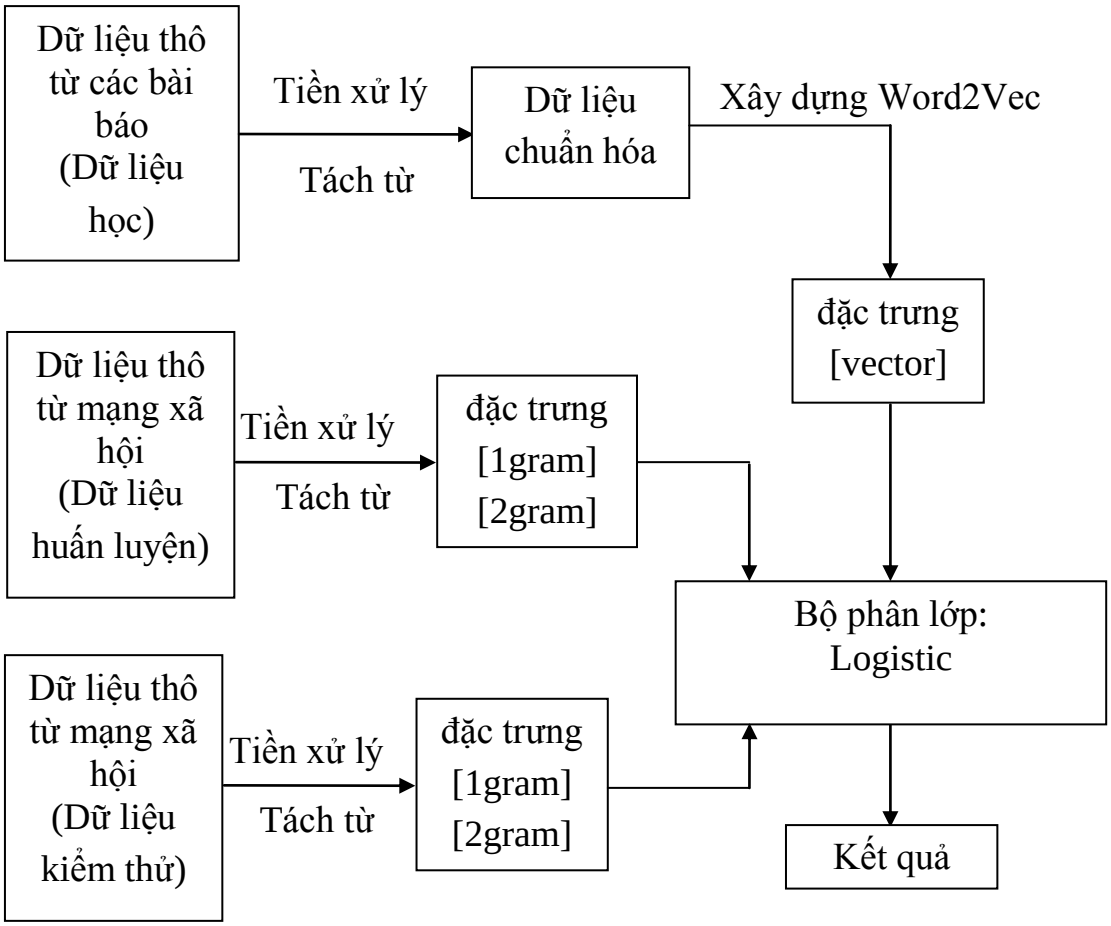
Từ	**Giá trị biểu diễn**
thuyền_thúng	0.7752657532691956
xe_đạp	0.7752000093460083
tàu_hỏa	0.749005913734436
xích_lô	0.7275688052177429
ô_tô	0.7588498592376709

Tiếp theo tôi thực hiện giống như với mô hình n-gram. Tuy nhiên khi tìm kiếm một liên kết giữa đặc trưng của dữ liệu kiểm thử với các đặc trưng trong tập huấn luyện. Nếu không tìm thấy bất kỳ một liên kết nào thì tôi sẽ đưa thêm Word2Vec vào bộ phân lớp để tìm kiếm. Tôi tìm từ của đặc trưng vector có giá trị biểu diễn gần nhất với từ của đặc trưng trong dữ liệu kiểm thử để sử dụng làm đặc trưng kiểm thử rồi tìm kiếm sự liên kết với các đặc trưng của tập huấn luyện. Cứ tiếp tục như vậy cho đến khi tìm được một liên kết với tập huấn luyện. Sau đó lại tiếp tục với các đặc trưng còn lại của dữ liệu kiểm thử. Cuối cùng giới tính của người dùng này sẽ được xác định dựa vào kết quả thu được.

[7] https://pypi.python.org/pypi/gensim

Ví dụ: Một đặc trưng của tập kiểm thử là: “đua xe_máy”, khi tìm kiếm trong đặc trưng của tập huấn luyện không có đặc trưng nào là “đua xe_máy” thì bộ phân lớp sẽ tìm trong Word2Vec thấy có từ “đua ô_tô” có giá trị biểu diễn gần bằng từ cần tìm. Khi đó bộ phân lớp sẽ dùng từ này làm đặc trưng để tìm kiếm trong đặc trưng của tập huấn luyện và thấy đặc trưng “đua ô_tô” nằm ở lớp giới tính nam nên bộ phân lớp sẽ cho kết quả là nam. Kết quả đặc trưng nằm ở lớp nào chiếm đa số sẽ được sử dụng làm kết quả cuối cùng.

Để dễ hình dung quy trình tôi có sơ đồ như sau:



Hình 3.2: Phân loại khi đưa thêm Word2Vec

3.3. Thực nghiệm

3.3.1. Dữ liệu thực nghiệm

* Chuẩn bị dữ liệu của người dùng mạng xã hội

- Sử dụng công cụ Graph API Explorer[8] để truy xuất vào và lấy nội dung trên tường của người dùng mạng xã hội.

[8] https://developers.facebook.com/tools/explorer

- Các nội dung trên tường của từng người được lưu vào 1 file theo định dạng:

<Blog>

<date> </date>

<post> </post>

</Blog>

- Sử dụng công cụ vn.vitk-master để tách từ, tách câu. Đối với dữ liệu tiếng Anh thì không phải thực hiện bước này còn đối với tiếng Việt thì bước này sẽ vô cùng cần thiết. VD: Tôi đi học bằng xe máy. Nếu không có bước này thì máy sẽ trích chọn đặc trưng từ 5 từ riêng lẻ sau: tôi, đi, học, bằng, xe, máy. Nếu tách từ thì máy sẽ trích chọn đặc trưng từ các từ như sau: tôi, đi_học, bằng, xe_máy. Như chúng ta thấy đối với hai trường hợp này thì việc tách từ sẽ đem đến độ chính xác cao hơn đối với phân loại văn bản tiếng Việt.

- Lưu tên file theo định dạng: mã.**giới tính**.tuổi.nghề nghiệp.tên.xml

Vd: 11106725783.**male**.25.student.tranvanchin.xml

Với mã, giới tính, tuổi, nghề nghiệp, tên là các nhãn được gán theo thông tin của người dùng mạng xã hội.

- Đưa tất cả các file vào một thư mục chung (khoảng 347 file)

VD: /home/trungkb/data/blogs

- Tạo ra và lưu vào 1 đối tượng SFrame[9] tất cả dữ liệu của người dùng mạng xã hội dưới dạng các đặc trưng và trọng số của các đặc trưng. Mỗi dòng là dữ liệu của một người dùng mạng xã hội.

Vd1: Chúng_mày ơi hôm_nay có đi_nhậu không?

Với ví dụ này thì thường các đặc trưng sẽ được trích chọn là: chúng_mày, đi_nhậu. Và các đặc trưng này thường là chỉ có người dùng có giới tính Nam hay sử dụng.

Vd2: Bạn nào đi mua_sắm với chúng_tớ không?

[9] https://github.com/turi-code/SFrame

Với ví dụ này thì thường các đặc trưng sẽ được trích chọn là: mua_sắm, chúng_tớ. Các đặc trưng này thể hiện người nói thường có giới tính Nữ.

Khi sử dụng Word2Vec thì nó sẽ tự học được thêm các từ có ngữ cảnh tương tự như các từ trên. Các từ có nghĩa gần giống nhau thì vector các từ này càng gần nhau. Ví dụ: chúng_mày, chúng_nó,... là các từ có nghĩa gần giống nhau nên vector các từ này rất gần nhau.

* Xây dựng Word2Vec để sử dụng làm 1 đặc trưng trong quá trình dự đoán.

- Tôi sử dụng thư viện Jsoup để viết mã lấy dữ liệu từ các trang web: 24h.com.vn, vnexpress.net, dantri.vn, eva.vn.

- Sau đó tôi sử dụng thêm công cụ vn.vitk-master để tách từ cho dữ liệu được lấy về.

- Lấy dữ liệu được sinh ra sau khi đã tách từ đưa hết vào thư mục /home/trungkb/data/txt.

- Sử dụng Gensim với tham số mặc định để sinh ra một file word2vec theo mô hình Skip-gram với số chiều của vector là 300, và bỏ qua các từ xuất hiện ít hơn 40 lần. Bộ dữ liệu để sinh ra file word2vec có dung lượng 1,2Gb bao gồm nội dung các bài báo được đăng trên các trang web như 24h.com.vn, vnexpress.net, dantri.vn, eva.vn,... Việc lấy nội dung này là tôi sử dụng thư viện Jsoup để lấy nội dung từ các trang web trên.

Khi xây dựng Word2Vec, dữ liệu đầu vào sẽ là một văn bản, xem như là tập hợp các từ (word). Đầu tiên, tương ứng với mỗi word thì chúng ta sẽ khởi tạo một vector ngẫu nhiên với số chiều được chỉ định (300 chiều). Sau khi đã có vector ngẫu nhiên, việc tiếp theo là thực hiện quá trình điều chỉnh vector của các từ này để sao cho chúng có thể biểu diễn được liên hệ giữa các từ có quan hệ với nhau.

Giả sử chúng ta có câu văn sau: ***Con mèo trèo cây cau.*** Tương ứng với mỗi từ trong câu này, chúng ta sẽ khởi tạo một vector ngẫu nhiên với số chiều được quy định trước (ví dụ số chiều = 50). Người ta sử dụng một mạng neuron và dùng mạng neural này để điều chỉnh dần dần các vector của các từ sao cho chúng thỏa mãn một số điều kiện nào đó. Câu hỏi đặt ra ở đây: Điều kiện đó là gì?

Để trả lời câu hỏi này thì trước hết chúng ta cần quan tâm tới một đặc điểm của ngôn ngữ, đó là những từ có mối liên hệ với nhau thường sẽ xuất hiện trong những ngữ cảnh khác nhau. Ví dụ từ "***trái***" và "***phải***" có thể xem là có mối liên quan nào đó với nhau vì nó đều dùng chỉ phương hướng và nó thường xuất hiện trong những mẫu câu giống nhau. Ví dụ tôi có các câu sau:

"Chạy xe phía bên trái", "Chạy ở bên phải", "Bên trái có vẻ rộng hơn", "Bên phải có một ngôi nhà".

Ta để ý thấy các từ nằm xung quanh của từ "***trái***" và "***phải***" đều khá là giống nhau không? Đó chính là nguyên tắc học của Word2Vec. Nó dựa vào những từ xung quanh của một từ nào đó để điều chỉnh vector của từ đó sao cho hợp lý.

Quay trở lại với ví dụ ban đầu: ***Con mèo trèo cây cau.*** Chúng ta sử dụng 1 mạng Neural để xem câu này có hợp lệ hay không. Giả sử thay từ "***trèo***" bằng từ "***ngủ***", rõ ràng chúng ta sẽ có 1 câu hoàn toàn vô nghĩa và hầu như không bao giờ xuất hiện trong văn bản bình thường: "***con mèo ngủ cây cau***". Bằng cách thay từ "***trèo***" bằng từ "***ngủ***" và nói cho mạng Neural biết rằng câu mới sinh ra là không hợp lệ, mạng Neural sẽ phải điều chỉnh các tham số trong mạng của nó một cách hợp lý để đưa ra được output đúng như chúng ta mong muốn (tức là "không hợp lệ"). Thông thường thì input vào mạng Neural sẽ không phải là nguyên một câu mà chỉ là 1 cụm từ của câu có độ dài dựa theo một tham số gọi là "window size". Ví dụ "window_size" = 3 thì chúng ta sẽ có các cụm từ: "***con mèo trèo***", "***mèo trèo cây***", "***trèo cây cau***". Với mỗi "windows size" thì chúng ta có thể thay 1 từ nào đó bằng 1 từ ngẫu nhiên khác để có các cụm câu vô nghĩa dùng để huấn luyện mạng Neural (bởi vì khi huấn luyện mạng Neural thì phải vừa cho đầu vào với nhãn "hợp lệ" và cũng phải có đầu vào với nhãn "không hợp lệ" nhằm giúp cho mạng Neural đó phân biệt cho đúng).

Nhờ việc huấn luyện mạng Neural trên một số lượng bài báo cực lớn tôi thu thập được từ các trang web ở trên thì vector của mỗi từ sẽ được điều chỉnh càng chính xác và những từ có liên quan nhau cũng sẽ xuất hiện ở gần nhau hơn. Khi đó giữa các từ sẽ có các mối liên hệ với nhau.

3.3.2. Cấu hình thực nghiệm

* Chuẩn bị môi trường

BeautifulSoup[10]: sử dụng để phân tích dữ liệu thô của người dùng mạng xã hội.

NLTK[11]: là một thư viện để xử lý loại bỏ các từ vô nghĩa, các ký tự không có nghĩa.

Gensim: là công cụ sử dụng để xây dựng Word2Vec.

GraphLab Create: sử dụng bộ phân lớp Logistic và đánh giá kết quả.

Python 2.7.6: viết mã sinh ra một đối tượng SFrame và xử lý dữ liệu để đưa vào huấn luyện Word2Vec.

JDK 1.7 và NetBeans IDE 8.0.2: viết mã để thu thập dữ liệu các bài báo từ các trang web 24h.com.vn, vnexpress.net, dantri.vn, eva.vn dùng để xây dựng Word2Vec và thu thập dữ liệu trên tường người dùng mạng xã hội để dùng làm dữ liệu huấn luyện và dữ liệu kiểm thử cho bài toán trên.

Tool: vn.vitk-master - xử lý tách từ trong tiếng Việt.

* Cấu hình máy tính thực nghiệm:

CPU Core I7

Ram 8Gb

HDD 250Gb

OS System: Ubuntu 14.10

3.3.3. Mô tả thực nghiệm

Đối với các tập dữ liệu trên tường người dùng mạng xã hội ở trên tôi lần lượt chia tập dữ liệu thành 2 phần có tỷ lệ như sau:

Bảng 3.2: Tỷ lệ chia tập dữ liệu huấn luyện và kiểm thử

Lần chia	*Tập huấn luyện*	*Tập kiểm thử*
1	*75%*	*25%*

[10] https://pypi.python.org/pypi/beautifulsoup4
[11] http://www.nltk.org/

2	*80%*	*20%*
3	*85%*	*15%*

Với lần chia thứ nhất tỷ lệ tập dữ liệu huấn luyện/tập dữ liệu kiểm thử là 75%-25% thì tôi sẽ thực nghiệm 10 lần. Mỗi lần thực nghiệm tôi sẽ lấy ngẫu nhiên dữ liệu theo tỷ lệ trên. Điều này sẽ giúp cho quá trình chạy thực nghiệm 10 lần thì cả 10 lần tập dữ liệu huấn luyện cũng như tập dữ liệu kiểm thử sẽ khác nhau. Sau đó tôi lấy trung bình cộng kết quả 10 lần chạy sẽ được kết quả thực nghiệm cho từng lần chia tỷ lệ tập dữ liệu.

Tương tự như trên với các lần chia tỷ lệ tập dữ liệu huấn luyện/tập dữ liệu kiểm thử là 80%-20% và 85%-15%. Với mỗi lần thực nghiệm tôi đều chạy theo cả 2 kỹ thuật thực hiện là: n-gram khi không có Word2Vec và khi có Word2Vec. Với mô hình n-gram tôi có các đặc trưng đầu vào là đặc trưng [1gram] và đặc trưng [2gram]. Khi sử dụng thêm Word2Vec tôi đưa thêm đặc trưng [vector] vào cho bộ phân lớp.

3.3.4. Đánh giá

Để đánh giá được độ hiệu quả của kỹ thuật phân loại giới tính người dùng dựa vào tin nhắn văn bản và Word2Vec, tôi tiến hành so sánh với một số kỹ thuật khác. Ở đây, tôi sẽ so sánh độ chính xác của việc sử dụng n-gram có Word2Vec so với khi không sử dụng Word2Vec.

Ta tiến hành xây dựng bộ phân lớp như sau:

Phân loại với đặc trưng [1gram] và [2gram]: Tiến hành huấn luyện bộ phân lớp với dữ liệu huấn luyện và dữ liệu kiểm thử là đặc trưng [1gram] và [2gram].

Phân loại khi sử dụng thêm Word2Vec: Tiến hành huấn luyện bộ phân lớp với dữ liệu huấn luyện là đặc trưng [1gram],[2gram] và [vector] còn dữ liệu kiểm thử là đặc trưng [1gram] và [2gram].

Tiêu chuẩn đánh giá trong thực nghiệm là độ đo chính xác, tỉ lệ phần trăm mẫu phân lớp chính xác trên tổng số mẫu kiểm thử, độ chính xác được tính bằng công thức sau:

$$\text{Độ chính xác} = \frac{|\{x|x \in D_{tst} \cap f(x) = y\}|}{|\{x|x \in D_{tst}\}|} \quad (3.1)$$

Trong đó, D_{tst} là dữ liệu kiểm thử, y là cực quan điểm ban đầu, $f(x)$ là cực quan điểm dự đoán.

3.3.5. Kết quả thực nghiệm

Kết quả thực nghiệm cho các trường hợp được nêu chi tiết dưới đây:

**** Tỷ lệ tập dữ liệu: 75% huấn luyện - 25% kiểm thử:***

Bảng 3.3: So sánh kết quả thực nghiệm với tỷ lệ tập dữ liệu 75%-25%

Lần thực nghiệm (75%-25%)	***[1gram]; [2gram]***	***[1gram]; [2gram];[vector]***
1	*0.538*	*0.817*
2	*0.559*	*0.839*
3	*0.548*	*0.849*
4	*0.538*	*0.817*
5	*0.548*	*0.882*
6	*0.570*	*0.839*
7	*0.538*	*0.613*
8	*0.559*	*0.527*
9	*0.581*	*0.806*
10	*0.581*	*0.838*
Trung bình	***0.556***	***0.783***

**** Tỷ lệ tập dữ liệu: 80% huấn luyện - 20% kiểm thử:***

Bảng 3.4: So sánh kết quả thực nghiệm với tỷ lệ tập dữ liệu 80%-20%

Lần thực nghiệm (80%-20%)	***[1gram]; [2gram]***	***[1gram]; [2gram];[vector]***
1	*0.608*	*0.838*
2	*0.595*	*0.527*
3	*0.608*	*0.865*
4	*0.608*	*0.838*
5	*0.608*	*0.878*
6	*0.608*	*0.838*
7	*0.595*	*0.824*
8	*0.608*	*0.824*
9	*0.608*	*0.851*
10	*0.622*	*0.878*
Trung bình	***0.607***	***0.816***

**** Tỷ lệ tập dữ liệu: 85% huấn luyện - 15% kiểm thử:***

Bảng 3.5: So sánh kết quả thực nghiệm với tỷ lệ tập dữ liệu 85%-15%

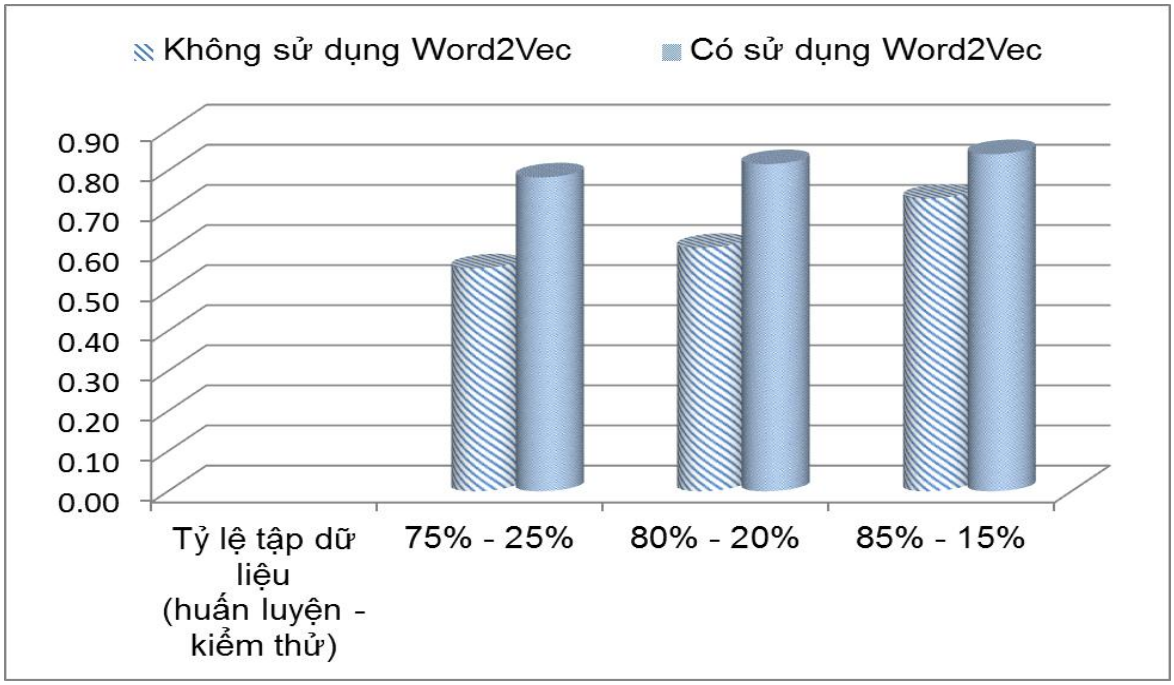
Lần thực nghiệm (85%-15%)	***[1gram]; [2gram]***	***[1gram]; [2gram];[vector]]***
1	*0.732*	*0.857*
2	*0.696*	*0.875*
3	*0.750*	*0.839*
4	*0.714*	*0.821*

5	*0.714*	*0.857*
6	*0.714*	*0.821*
7	*0.768*	*0.839*
8	*0.732*	*0.821*
9	*0.732*	*0.839*
10	*0.750*	*0.839*
Trung bình	***0.730***	***0.841***

Bảng 3.6: Tổng hợp so sánh kết quả thực nghiệm

Tỷ lệ tập dữ liệu (huấn luyện - kiểm thử)	***[1gram]; [2gram]***	***[1gram]; [2gram];[vector]***
75% - 25%	*0.556*	*0.783*
80% - 20%	*0.607*	*0.816*
85% - 15%	*0.730*	*0.841*

Để thấy rõ và dễ so sánh kết quả thực nghiệm hơn ta sẽ xem biểu đồ sau:



Hình 3.3: Biểu đồ biểu diễn kết quả thực nghiệm

Nhận xét: Dựa vào bảng tổng hợp kết quả trên và biểu đồ thì ta thấy đối với bài toán này phân loại giới tính người dùng dựa vào tin nhắn văn bản khi tăng dữ liệu tập huấn luyện lên càng cao thì kết quả thu về càng chính xác. Bên cạnh đó khi sử dụng thêm Word2Vec để làm đặc trưng cho n-gram thì kết quả thu được luôn tốt hơn khi không sử dụng Word2Vec.

KẾT LUẬN

Với sự phát triển không ngừng của mạng xã hội như ngày nay, nhu cầu phân tích và tìm hiểu thông tin về người dùng là rất lớn. Các công ty rất muốn có những thông tin về người dùng để phục vụ cho mục đích kinh doanh của họ. Chính vì vậy mà việc phân loại được các thông tin người dùng một cách tự động là việc làm hết sức thiết thực hiện nay. Đối với luận văn này tôi đã trình bày phương pháp để có thể phân loại giới tính người dùng mạng xã hội một cách tự động.

Trong quá trình thực hiện luận văn, tôi đã trình bày một số khái niệm cơ bản về khai phá dữ liệu, quá trình khai phá dữ liệu, một số chức năng chính của khai phá dữ liệu cùng với một số kỹ thuật khai phá dữ liệu hiện nay. Ngoài ra tôi còn trình bày một số khái niệm cũng như những lợi ích và tác hại của mạng xã hội đối với cuộc sống của chúng ta hiện nay. Bên cạnh đó tôi cũng giới thiệu một số mạng xã hội phổ biến.

Về mặt phương pháp luận tôi đã giới thiệu tổng quan về Word2Vec và mô hình từ thành vector: vector từ, lập luận với vector từ và nghiên cứu về vector từ. Cùng với đó tôi giới thiệu các mô hình Continuous Bag-of-word và Skip-gram được đề xuất bởi Tomas Mikolov và cộng sự nhằm giải thích rõ hơn cách biểu diễn các từ dưới dạng Word2Vec.

Về thực nghiệm, tôi đã sử dụng thư viện Jsoup và viết code Java tự thu thập và tiền xử lý dữ liệu để xây dựng đặc trưng Word2Vec từ các bài báo trên các trang web. Bên cạnh đó tôi đã tiền xử lý và xây dựng các đặc trưng cho bộ dữ liệu huấn luyện từ dữ liệu tự thu thập trên tường người dùng mạng xã hội bằng thư viện Graph API Explorer. Do đều là các dữ liệu Tiếng Việt nên trước khi sử dụng tôi đều phải sử dụng công cụ vn.vitk-master để tách từ. Sau đó tôi thực nghiệm với các tỷ lệ dữ liệu khác nhau sử dụng mô hình phân loại n-gram khi không sử dụng Word2Vec và khi có sử dụng Word2Vec. Sau đó tôi sử dụng độ đo từ các kết quả thu được và chứng minh được khi sử dụng mô hình phân loại n-gram với việc sử dụng thêm Word2Vec kết quả đạt được là tốt hơn.

Hướng phát triển

Do sự nhập nhằng của dữ liệu Tiếng Việt cũng như kiến thức của bản thân còn hạn chế nên kết quả thực nghiệm cho Tiếng Việt còn chưa cao như mong muốn. Tôi cần phải cải tiến phương pháp và xử lý dữ liệu tốt hơn để đạt được

hiệu quả cao hơn nữa. Bên cạnh đó tôi sẽ thử nghiệm việc phân loại trên các thuộc tính khác nữa của người dùng mạng xã hội như: độ tuổi, sở thích, ... Sau khi có được kết quả thực nghiệm như mong muốn tôi sẽ nghiên cứu xây dựng một hệ thống tự động hóa việc dự đoán thông tin người dùng mạng xã hội.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt:

[1] Nguyễn Thị Thanh Thảo, ***Tìm hiểu các ứng dụng của datamining trong kinh doanh***, 2012.

[2] Nhóm tác giả: Kim Đình Sơn, Đặng Ngọc Thuyên, Phùng Văn Chiến, Ngô Thành Đạt, ***Các mô hình ngôn ngữ N-gram và Ứng dụng***, 2013.

[3] Bộ môn hệ thống thông tin, Khoa công nghệ thông tin, Đại học hàng hải Việt Nam, ***Bài giảng khai phá dữ liệu***, 2011.

[4] Bộ phận tư vấn – hỗ trợ và giới thiệu việc làm SV, ***Tác động của mạng xã hội đến học sinh sinh viên***, 2015.

https://www.kgtec.edu.vn/component/k2/1440-tac-dong-cua-mang-xa-hoi-den-hoc-sinh-sinh-vien.

Tài liệu tiếng Anh:

[5] Andriy Mnih and Geoffrey E Hinton. ***A scalable hierarchical distributed language model. Advances in neural information processing systems, 21:1081–1088***, 2009.

[6] Andriy Mnih and Yee Whye Teh. ***A fast and simple algorithm for training neural probabilistic language models. arXiv preprint arXiv:1206.6426***, 2012.

[7] David A. Jurgens, Saif M. Mohammad, Peter D. Turney, Keith J. Holyoak, ***SemEval-2012 Task 2: Measuring Degrees of Relational Similarity***, 2012.

[8] Frederic Morin and Yoshua Bengio. ***Hierarchical probabilistic neural network language model. In Proceedings of the international workshop on artificial intelligence and statistics, pages 246–252***, 2005.

[9] Michael U Gutmann and Aapo Hyv¨arinen. ***Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. The Journal of Machine Learning Research, 13:307–361***, 2012.

[10] Mikolov et al, ***Distributed Representations of Words and Phrases and their Compositionality***, 2013.

[11] Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. ***Efficient estimation of word representations in vector space. ICLR Workshop***, 2013.

[12] Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. ***Extensions of recurrent neural network language model. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5528–5531. IEEE***, 2011.

[13] Richard Socher, Yoshua Bengio and Chris Manning, ***Deep Learning for NLP (without Magic)***, ACL2012.

[14] Ronan Collobert and Jason Weston. ***A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167. ACM***, 2008.

[15] Rong, ***Word2vec Parameter Learning Explained***, 2014.

[16] Margaret Rouse, ***Social networking***, 2016.

http://whatis.techtarget.com/definition/social-networking.