HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN BÁ QUYỀN

NGHIÊN CỨU PHÂN TÍCH HÀNH VI MUA HÀNG CỦA KHÁCH HÀNG SỬ DỤNG MẠNG NƠRON

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2021

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN BÁ QUYỀN

NGHIÊN CỨU PHÂN TÍCH HÀNH VI MUA HÀNG CỦA KHÁCH HÀNG SỬ DỤNG MẠNG NƠRON

CHUYÊN NGÀNH : KHOA HỌC MÁY TÍNH

MÃ SỐ: 8.48.01.01

ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. PHẠM HOÀNG DUY

HÀ NỘI – 2021

LỜI CAM ĐOAN

Tôi xin cam đoan những nội dung trong luận văn này là do tôi thực hiện dưới sự hướng dẫn của **TS. Phạm Hoàng Duy**. Mọi tham khảo dùng trong luận văn đều được trích dẫn nguồn gốc rõ ràng. Các nội dung nghiên cứu và kết quả trong đề tài này là trung thực và chưa từng được ai công bố trong bất cứ công trình nào.

TÁC GIẢ

Nguyễn Bá Quyền

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

THUẬT NGỮ VIẾT TẮT

Kí hiệu	Ý nghĩa
RNN(Recurrent Nơron network)	Mạng Nơron hồi quy
LSTM(Long short-term memory)	Kiến trúc trí nhớ ngắn hạn, dài hạn
RNN-Emb-Word2vec	Phép nhúng kết hợp với Word2vec.
RNN-Emb-Jointly-Lin	Phép nhúng kết hợp phân loại tuyến tính.
RNN-Emb-Jointly-Nonlin	Phép nhúng kết hợp phân loại phi tuyến tính.
RNN-Emb-Word2vec-Finetune	Phép nhúng riêng biệt và được điều chỉnh vi cấp chung
RNN-Emb-Output	Phép nhúng được đào tạo riêng biệt
RNN-Att-HS-Lin	Cơ chế chú ý đến các trạng thái RNN ẩn với trọng số chú ý tuyến tính
RNN-Att-HS-Nonlin	Cơ chế chú ý đến các trạng thái RNN ẩn với trọng số chú ý phi tuyến tính
RNN-Att-Emb-Lin	Cơ chế chú ý trỏ đến các phép nhúng có trọng số tuyến tính
RNN-Att-Emb-Nonlin	Cơ chế chú ý trỏ đến các phép nhúng có trọng số phi tuyến tính

MỞ ĐẦU

Dự đoán hành vi của khách hàng trong tương lai là một nhiệm vụ quan trọng để mang lại cho họ trải nghiệm tốt nhất có thể và cải thiện sự hài lòng của họ. Một ví dụ thực tế được quan sát thấy trong các hệ thống thương mại điện tử, nơi khách hàng có thể tránh việc tìm kiếm thông qua một danh mục sản phẩm không thực sự cần thiết và thay vào đó họ có một bộ sản phẩm được đề xuất đáp ứng được điều họ quan tâm. Hành vi của khách hàng có thể được biểu diễn dưới dạng dữ liệu tuần tự mô tả các tương tác qua thời gian, ví dụ về những tương tác này là các mặt hàng mà khách hàng mua hoặc xem. Do đó, lịch sử tương tác của khách hàng có thể được mô hình hóa dưới dạng dữ liệu tuần tự có đặc điểm cụ thể và có thể kết hợp với nhau thông qua khía cạnh thời gian. Để kiểm tra, nếu khách hàng mua điện thoại di động mới, họ có thể mua phụ kiện cho điện thoại di động này trong tương lai gần hoặc nếu khách hàng mua sách, họ có thể quan tâm đến sách của cùng tác giả hoặc thể loại. Để đưa ra dự đoán chính xác là điều rất quan trọng. Một cách phổ biến để xử lý dữ liệu này là xây dựng các tính năng thủ công để tổng hợp thông tin từ các dữ liệu trong quá khứ. Ví dụ: người ta có thể đếm số lượng sản phẩm đã mua của một danh mục cụ thể trong N ngày qua hoặc số ngày kể từ lần mua cuối cùng. Việc tạo một số tính năng được làm thủ công sẽ tạo ra một vectơ đặc trưng có thể được đưa vào một thuật toán học máy như hồi quy logistic. Mặc dù có thể đạt được kết quả tốt với phương pháp này, nhưng nó có một số mặt hạn chế. Đầu tiên, một phần của mối quan hệ giữa thời gian và trình tự bị bỏ qua. Mặc dù chúng có bao gồm các tính năng chứa thông tin từ các tương tác trong quá khứ nhưng trên thực tế vẫn có thể bao gồm tất cả thông tin có trong dữ liệu thô. Chỉ các tín hiệu được mã hóa trong các tính năng này mới có thể được các mô hình dự đoán ghi lại. Thứ hai, thông thường sẽ có một tập hợp rất lớn các tính năng được tạo thủ công bằng tay. Các nhà khoa học dữ liệu có thể dành nhiều thời gian để thiết kế và thử nghiệm các tính năng mới, mà nhiều tính năng trong số đó không cải thiện hiệu suất dự đoán. Ngay cả khi họ có thể cải thiện, rất khó để biết liệu tập hợp các tính năng thủ công trên thực tế có tối ưu cho vấn đề hay không, vì vậy quá trình thử nghiệm và thêm các

tính năng thủ công mới không bao giờ dừng lại hoặc chỉ dừng lại khi thuật toán đạt mức chấp nhận được. Mức hiệu suất có thể khác xa so với tiềm năng thực sự. Thứ ba, trong một số trường hợp, việc tính toán các tính năng thủ công có thể dẫn đến việc xử lý dữ liệu khá tốn kém.

Với việc học sâu nhận được rất nhiều sự chú ý trong những năm qua, một cách tiếp cận mới đối với dữ liệu tuần tự đã được khám phá. Mạng thần kinh tái tạo (RNN) rất năng động để học các mẫu tuần tự phức tạp, vì chúng có khả năng duy trì trạng thái ẩn được cập nhật bởi một hàm phi tuyến tính phức tạp được học từ chính dữ liệu. Họ có thể nắm bắt thông tin về sự phát triển của những gì đã xảy ra trong các bước thời gian trước đó. Trong những năm qua, RNN đã đạt được trình độ tiên tiến trong các vấn đề như lập mô hình ngôn ngữ, nhận dạng giọng nói, dịch máy hoặc nhận dạng chữ viết tay, các tác vụ này có một số điểm tương đồng với bài toán dự đoán các hành động trong tương lai từ dữ liệu tương tác trong quá khứ, theo nghĩa là dữ liệu được biểu diễn tuần tự.

Luận văn này là nghiên cứu các kỹ thuật khác nhau khi sử dụng RNN để dự đoán hành vi của khách hàng trong tương lai. Cụ thể hơn, tập trung vào hai khía cạnh: Nghiên cứu các phép nhúng có thể được sử dụng để tạo ra các đại diện mục vectơ hữu ích giúp cải thiện các dự đoán với RNN. Tiếp theo đánh giá và phân tích các biểu diễn vectơ của các lựa chọn thay thế khác nhau để tìm hiểu mục nhúng. Nghiên cứu cách các cơ chế chú ý có thể giúp giải thích các dự đoán của các mô hình RNN. Sau đó phân tích hiệu suất của các biến thể cơ chế chú ý khác nhau và cung cấp các ví dụ trong đó các dự đoán được giải thích bằng các nguyên tắc trong quá khứ.

CHƯƠNG 1. PHÂN TÍCH HÀNH VI KHÁCH HÀNG

Trong chương này, chúng ta tập trung xác định các thành phần của quá trình ra quyết định mà khách hàng tuân theo, các mô hình chính được giới thiệu trong nghiên cứu hành vi khách hàng, nghiên cứu về lý thuyết cho việc đưa ra quyết định. Ngoài ta, trong chương này, tập trung vào việc xác định vấn đề cần nghiên cứu và phạm vi nghiên cứu. Và nêu những khó khăn gặp phải trong quá trình nghiên cứu và cách khắc phục.

1.1. Khái quát về hành vi khách hàng.

1.1.1. *Giới thiệu chung.*

Bất kỳ người nào tham gia vào quá trình tiêu dùng đều là khách hàng. Khách hàng là cá nhân thực hiện việc mua để phục vụ tiêu dùng cá nhân hoặc để đáp ứng nhu cầu tập thể của gia đình và nhu cầu hộ gia đình. Hành vi của khách hàng có nghĩa là cách các cá nhân đưa ra quyết định sử dụng các nguồn lực sẵn có của họ như thời gian, tiền bạc, nỗ lực để tiêu dùng các sản phẩm và dịch vụ khác nhau. Nó bao gồm những gì họ mua, tại sao họ mua nó, khi họ mua nó, họ mua nó ở đâu, tần suất mua nó và tần suất sử dụng nó. Hành vi của khách hàng là những hành động mà một người thực hiện trong việc mua và sử dụng các sản phẩm và dịch vụ, bao gồm các quá trình tinh thần và xã hội diễn ra trước và sau các hành động này. Leon G.Schiffman và Leslie lazar Kanuk đã định nghĩa hành vi của khách hàng "là hành vi mà khách hàng thể hiện khi tìm kiếm, mua, sử dụng, đánh giá và loại bỏ các sản phẩm, dịch vụ và ý tưởng mà họ mong đợi sẽ thỏa mãn nhu cầu của họ". Hành vi của khách hàng là cách một cá nhân hành động trong khi thu nhận và sử dụng hàng hóa và dịch vụ. Hành động này liên quan đến một quá trình quyết định, cá nhân bị ảnh hưởng bởi các đặc điểm cá nhân và các yếu tố môi trường. Hành vi của khách hàng là một chủ đề rộng lớn và phức tạp. Hiểu được hành vi của khách hàng và "hiểu biết về khách hàng" không hề đơn giản. Hầu như không thể dự đoán chính xác, khách hàng sẽ hành xử như thế nào trong một tình huống nhất định. Những nỗ lực của tất cả các nhà tiếp thị là để tác động đến hành vi của khách hàng theo cách thức họ mong muốn. Sự thành công hay thất bại trong việc theo đuổi này quyết định

sự khác biệt giữa thành công và thất bại của các nỗ lực tiếp thị hoặc thậm chí của chính doanh nghiệp.

Cần nghiên cứu Hành vi của Khách hàng:

Các nhà tiếp thị phải có quyền truy cập vào dữ liệu liên quan đến khách hàng, thói quen mua hàng và loại phương tiện truyền thông nào mà họ ưa thích, để phát triển các chương trình truyền thông thuyết phục.

- Ai là người đưa ra quyết định mua?
- Ai là người ảnh hưởng đến quyết định mua?
- Điều gì thúc đẩy người mua và mọi người thực hiện hành động?

1.1.2. *Các yếu tố ảnh hưởng đến hành vi tiêu dùng*

Việc mua hàng của khách hàng bị ảnh hưởng mạnh mẽ bởi các đặc điểm văn hóa, xã hội, cá nhân và tâm lý.

1.1.2.1. *Yếu tố văn hóa.*

Văn hóa: Văn hóa là tập hợp các giá trị, nhận thức, mong muốn và hành vi cơ bản mà một thành viên trong xã hội học được từ gia đình và các thiết chế quan trọng khác. Về cơ bản, văn hóa là một phần của mọi xã hội và là nguyên nhân quan trọng của mong muốn và hành vi cá nhân. Ảnh hưởng của văn hóa đến hành vi mua hàng ở mỗi quốc gia khác nhau, do đó các nhà tiếp thị phải rất cẩn thận trong việc phân tích văn hóa của các nhóm, khu vực hoặc thậm chí các quốc gia khác nhau.

Văn hóa phụ: Mỗi nền văn hóa chứa các nền văn hóa con khác nhau như tôn giáo, dân tộc, vùng địa lý, nhóm chủng tộc, v.v. Các nhà tiếp thị có thể sử dụng các nhóm này bằng cách phân khúc thị trường thành nhiều phần nhỏ khác nhau. Các nhà tiếp thị có thể thiết kế sản phẩm theo nhu cầu của một nhóm địa lý cụ thể.

Giai cấp xã hội: Giai cấp xã hội đề cập đến sự sắp xếp thứ bậc của xã hội thành nhiều bộ phận khác nhau, mỗi bộ phận biểu thị địa vị hoặc vị thế xã hội. Tầng lớp xã hội là một yếu tố quan trọng quyết định hành vi của khách hàng vì nó ảnh hưởng đến mô hình tiêu dùng, lối sống, mô hình phương tiện truyền thông, hoạt động và lợi ích của khách hàng.

1.1.2.2. *Yếu tố xã hội*

Các yếu tố xã hội cũng tác động đến hành vi mua của khách hàng. Các yếu tố xã hội quan trọng là: nhóm tham chiếu, gia đình, vai trò và địa vị.

Nhóm tham chiếu: Nhóm tham chiếu người là những nhóm có ảnh hưởng trực tiếp hoặc gián tiếp đến thái độ hoặc hành vi của người đó. Các cá nhân sử dụng các nhóm này làm điểm tham chiếu để học hỏi thái độ, niềm tin và hành vi, và thích ứng với những điều này trong cuộc sống của họ. Gia đình và bạn thân được coi là những nhóm tham chiếu chính trong cuộc sống của một cá nhân do tần suất tương tác của họ với cá nhân và mức độ quan trọng của những người khác quan trọng này trong cuộc sống của một cá nhân. Bạn cùng trường, khu phố, đồng nghiệp, những người quen khác là một phần của nhóm tham chiếu thứ cấp của một cá nhân.

Gia đình: Hành vi của người mua bị ảnh hưởng mạnh mẽ bởi các thành viên trong một gia đình. Do đó các nhà tiếp thị đang cố gắng tìm ra vai trò và ảnh hưởng của người chồng, người vợ và con cái. Nếu quyết định mua một sản phẩm cụ thể bị ảnh hưởng bởi vợ thì các nhà tiếp thị sẽ cố gắng nhắm mục tiêu vào phụ nữ trong quảng cáo của họ. Ở đây, chúng ta cần lưu ý rằng vai trò mua thay đổi cùng với sự thay đổi trong lối sống của khách hàng.

Vai trò và Trạng thái: Mỗi người có các vai trò và địa vị khác nhau trong xã hội tùy thuộc vào các nhóm, câu lạc bộ, gia đình, tổ chức, v.v. mà anh ta thuộc về. Vai trò và địa vị xã hội ảnh hưởng sâu sắc đến hành vi của khách hàng và quyết định mua hàng của anh ta.

1.1.2.3. *Yếu tố cá nhân.*

Các yếu tố cá nhân cũng có thể ảnh hưởng đến hành vi của khách hàng. Một số yếu tố cá nhân quan trọng ảnh hưởng đến hành vi mua là: lối sống, hoàn cảnh kinh tế, nghề nghiệp, tuổi tác, tính cách và quan niệm về bản thân.

Tuổi tác: Tuổi và vòng đời có tác động tiềm tàng đến hành vi mua của khách hàng. Khách hàng thay đổi việc mua hàng hóa và dịch vụ theo thời gian. Vòng đời gia đình bao gồm các giai đoạn khác nhau như thời thơ ấu, giai đoạn độc thân, vợ chồng mới cưới, làm cha mẹ, ... giúp các nhà tiếp thị phát triển các sản phẩm phù

hợp cho từng giai đoạn.

Nghề nghiệp: Nghề nghiệp của một người có tác động đáng kể đến hành vi mua hàng của anh ta. Ví dụ, một giám đốc tiếp thị của một tổ chức sẽ cố gắng mua những bộ quần áo công sở, trong khi một nhân viên cấp thấp trong cùng một tổ chức sẽ mua những bộ quần áo bảo hộ lao động thô kệch.

Tình hình kinh tế: Tình hình kinh tế của khách hàng có ảnh hưởng lớn đến hành vi mua hàng của anh ta. Nếu thu nhập và tiết kiệm của khách hàng cao thì anh ta sẽ mua những sản phẩm đắt tiền hơn. Mặt khác, một người có thu nhập thấp và tiết kiệm sẽ mua những sản phẩm rẻ tiền.

Cách sống: Phong cách sống của khách hàng là một yếu tố nhập khẩu khác ảnh hưởng đến hành vi mua của khách hàng. Phong cách sống đề cập đến cách một người sống trong một xã hội và được thể hiện bằng những thứ xung quanh họ. Nó được xác định bởi sở thích, ý kiến, hoạt động của khách hàng, v.v. và định hình toàn bộ khuôn mẫu hành động và tương tác của anh ta trên thế giới.

Nhân cách: Tính cách thay đổi từ người này sang người khác, thời gian và địa điểm. Do đó nó có thể ảnh hưởng rất lớn đến hành vi mua của khách hàng. Trên thực tế, Nhân cách không phải là những gì người ta mặc; đúng hơn nó là tổng thể các hành vi của một người đàn ông trong những hoàn cảnh khác nhau. Nó có các đặc điểm khác nhau như: thống trị, hiếu chiến, tự tin, v.v. có thể hữu ích để xác định hành vi của khách hàng đối với sản phẩm hoặc dịch vụ cụ thể.

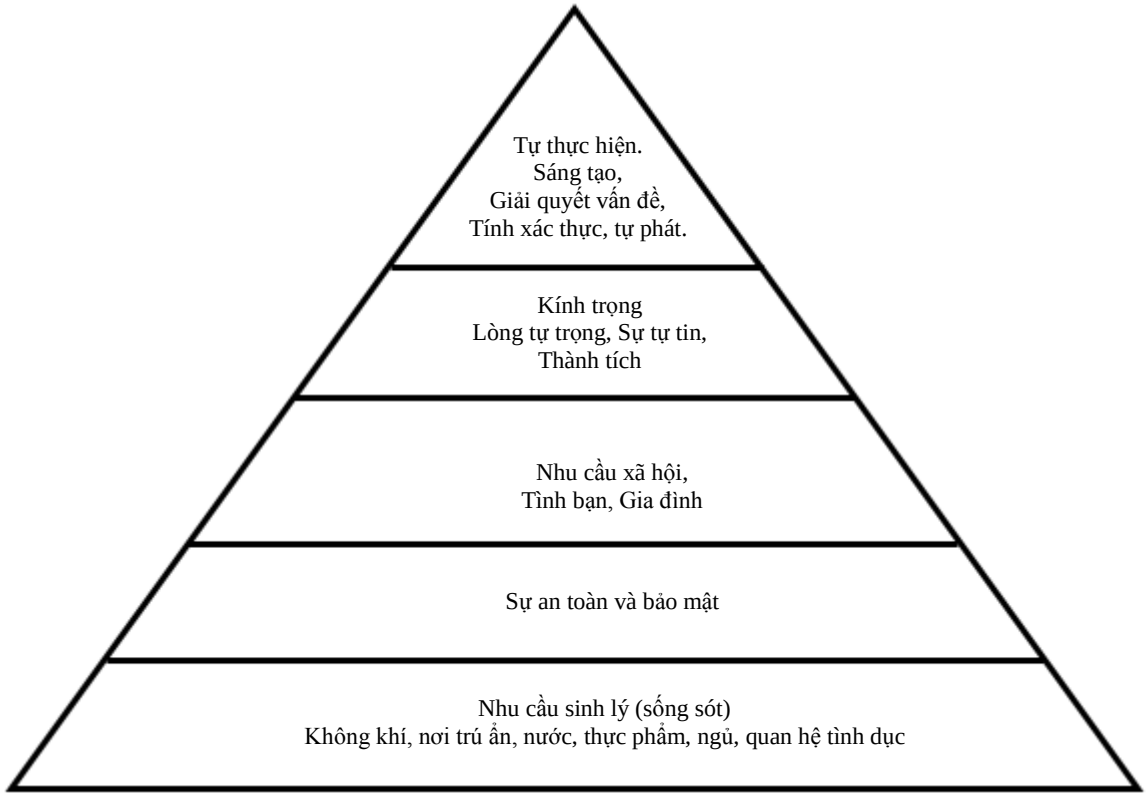
1.1.2.4. *Yếu tố tâm lý.*

Có bốn yếu tố tâm lý quan trọng ảnh hưởng đến hành vi mua của khách hàng. Đó là: nhận thức,động cơ, học tập, niềm tin và thái độ.

Động lực: Mức độ của động cơ cũng ảnh hưởng đến hành vi mua của khách hàng. Mỗi người đều có những nhu cầu khác nhau như nhu cầu sinh lý, nhu cầu sinh học, nhu cầu xã hội ... Bản chất của các nhu cầu là ở chỗ, một số nhu cầu bức xúc nhất trong khi một số khác lại ít bức xúc nhất. Do đó, nhu cầu trở thành động cơ khi càng thúc ép người đó tìm kiếm sự thỏa mãn.

Lý thuyết về Động lực của Maslow giải thích lý do tại sao mọi người bị thúc đẩy

bởi những nhu cầu cụ thể vào những thời điểm cụ thể. Maslow đã sắp xếp các nhu cầu của con người theo thứ bậc theo mức độ quan trọng của chúng. Đó là nhu cầu sinh lý, nhu cầu an toàn, nhu cầu xã hội, nhu cầu về lòng tự trọng và nhu cầu hiện thực hóa bản thân. Một người cố gắng thỏa mãn nhu cầu quan trọng nhất trước tiên. Khi nhu cầu đó được thỏa mãn, nó sẽ không còn là động lực nữa và người đó sau đó sẽ cố gắng thỏa mãn nhu cầu quan trọng tiếp theo.

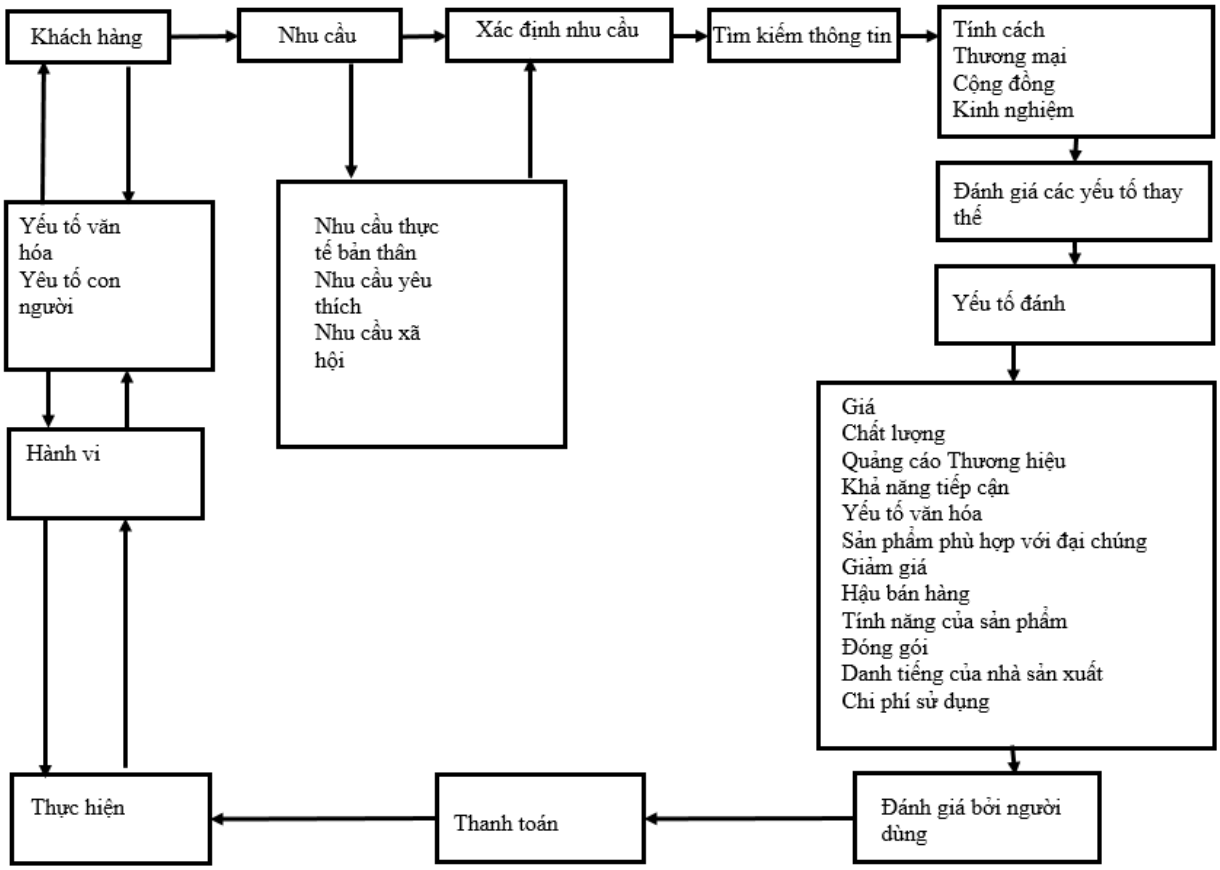


Hình 1.1. Lý thuyết về động lực của Maslow

Nhận thức: Lựa chọn, sắp xếp và giải thích thông tin theo cách để tạo ra trải nghiệm có ý nghĩa về thế giới được gọi là nhận thức. Những gì một cá nhân nghĩ về một sản phẩm hoặc dịch vụ cụ thể là nhận thức của họ đối với cùng một. Những người có cùng nhu cầu có thể không mua các sản phẩm tương tự do sự khác biệt về nhận thức. Có ba quá trình tri giác khác nhau là sự chú ý có chọn lọc, sự bóp méo có chọn lọc và sự duy trì có chọn lọc. Trong trường hợp không chú ý, các cá nhân chú ý đến thông tin được sử dụng cho họ hoặc các thành viên gia đình trực tiếp của họ. Trong khi đó, trong trường hợp bóp méo có chọn lọc, khách hàng có xu hướng nhận thức thông tin theo cách phù hợp với suy nghĩ và niềm tin hiện có của họ.

Tương tự, trong trường hợp giữ chân có chọn lọc, khách hàng ghi nhớ thông tin hữu ích cho họ,

Niềm tin và thái độ: Khách hàng có niềm tin và thái độ cụ thể đối với các sản phẩm khác nhau. Vì niềm tin và thái độ như vậy tạo nên hình ảnh thương hiệu và ảnh hưởng đến hành vi mua hàng của khách hàng, do đó, các nhà tiếp thị quan tâm đến chúng. Các nhà tiếp thị có thể thay đổi niềm tin và thái độ của khách hàng bằng cách tung ra các chiến dịch đặc biệt về vấn đề này.



Hình 1.2. Mô hình khái niệm cho Quy trình quyết định mua của khách hàng & hành vi của khách hàng.

Mô hình này giải thích quá trình quyết định mua hàng của khách hàng và hành vi của khách hàng. Bước đầu tiên là xác định nhu cầu chưa được thỏa mãn. Sau đó, thông tin có thể được tìm kiếm từ các nguồn khác nhau như nguồn cá nhân, thương mại, công cộng và kinh nghiệm. Sau khi hoàn thành quá trình tìm kiếm, khách hàng sẽ nhận được rất nhiều Khách hàng đánh giá các lựa chọn có sẵn bằng cách sử dụng các tiêu chí đánh giá nhất định. chúng là giá cả, chất lượng, quảng cáo, thương hiệu,

v.v. Sau khi đánh giá, việc mua hàng thực tế sẽ diễn ra nhanh chóng. Cuối cùng, giai đoạn quan trọng là sau khi quyết định mua hàng nghĩa là nếu khách hàng hài lòng với sản phẩm, họ sẽ tiếp tục mua sản phẩm đó, nếu không họ sẽ chuyển sang sản phẩm khác. Quá trình quyết định mua hàng của thương mại là một quá trình liên tục.

1.1.3. *Đóng góp của luận văn và các kỹ thuật liên quan.*

Đóng góp chính của luận văn này là nghiên cứu các kỹ thuật khác nhau khi sử dụng RNN để dự đoán hành vi của khách hàng trong tương lai. Cụ thể hơn, luận văn tập trung vào hai khía cạnh:

- Nghiên cứu phép nhúng có thể được sử dụng để tạo ra các biểu diễn mục vectơ hữu ích giúp cải thiện các dự đoán với RNN và đánh giá, phân tích biểu diễn vectơ của các lựa chọn thay thế khác nhau để tìm hiểu cách nhúng mục.
- Nghiên cứu cách các cơ chế chú ý có thể giúp giải thích các dự đoán về Các mô hình RNN và phân tích hiệu suất của các cơ chế chú ý khác nhau cách biến thể và cung cấp các ví dụ trong đó các dự đoán được giải thích bằng các tương tác trong quá khứ.

Để thực việc nghiên cứu, luận văn tập trung vào việc nghiên cứu các phép nhúng và cơ chế chú ý.

1.2. Phạm vi công việc nghiên cứu.

1.2.1. *Khái quát vấn đề.*

Phân khúc thị trường là một quá trình đòi hỏi xác định các loại nhóm khách hàng đồng nhất được mô tả bởi một tập hợp các đặc điểm tương đồng, để cải thiện các hoạt động tiếp thị thông qua việc phân bổ nguồn lực và xây dựng chiến lược tùy biến tốt hơn. Khi các nhóm mục tiêu được biết đến trước, vấn đề sẽ trở thành một nhiệm vụ phân loại, theo một quá trình học tập có giám sát. Sự quan tâm ngày càng tăng trong việc xác định các nguồn khách hàng mới, buộc các tổ chức tài chính phải điều tra các phương pháp mới để phát hiện các cá nhân có xu hướng tiết kiệm tiền cao, và sử dụng trong việc chi tiêu. Các phương pháp thống kê truyền thống như

phân tích phân biệt thường được sử dụng trong các nhiệm vụ phân loại, mang lại kết quả tốt. Tuy nhiên, nhu cầu thu được kết quả chính xác hơn nữa đã khiến các nhà nghiên cứu quan tâm đến các kỹ thuật phân loại không tham số như mạng lưới thần kinh nhân tạo. Mục đích chính của nghiên cứu này là phân tích kết quả thu được khi xây dựng mô hình xác định các cá nhân có cơ hội lớn để thực hiện việc mua hàng, sử dụng mạng lưới thần kinh nhân tạo.

1.2.2. *Mục đích nghiên cứu.*

Luận văn này nghiên cứu phép nhúng được sử dụng để tạo ra các đại diện mục vectơ hữu ích giúp cải thiện các dự đoán với RNN. Luận văn sẽ trình bày việc đánh giá và phân tích các biểu diễn vectơ của các lựa chọn thay thế khác nhau để tìm hiểu cách nhúng mục. Ngoài ta, trong luận văn còn nghiên cứu cách các cơ chế chú ý có thể giúp giải thích các dự đoán của các mô hình RNN. Thêm nữa, luận văn này còn trình bày việc phân tích hiệu suất của các biến thể cơ chế chú ý khác nhau và cung cấp các ví dụ trong đó các dự đoán được giải thích bằng các nguyên tắc trong quá khứ.

1.2.3. *Đối tượng và phạm vi nghiên cứu.*

Luận văn tập trung vào nghiên cứu bài toán phân tích hành vi khách hàng tập trung vào hành vi mua hàng nhằm xác định các cá nhân có cơ hội lớn để thực hiện việc mua hàng, sử dụng mạng lưới thần kinh nhân tạo.

Bộ dữ liệu phân tích

Tập dữ liệu được sử dụng là tập dữ liệu danh sách các bộ phim được bán ra và được đánh giá bởi khách hàng từ năm 2008 đến năm 2015[1]. Tập dữ liệu bao gồm lịch sử xếp hạng các bộ phim do khách hàng khác nhau đánh giá. Bộ dữ liệu xếp hạng chứa một giá trị là mốc thời gian, cho biết thứ tự mà khách hàng xếp hạng các bộ phim. Trong tập dữ liệu này, các bộ phim được xếp hạng theo điểm số từ 1 đến 5. Tuy nhiên, trong luận văn này chỉ sử dụng các dữ liệu đã khách hàng đã xếp hạng một bộ phim chứ không phải điểm số. Ví dụ về lịch sử tương tác của khách hàng có thể được mô tả như sau:

• 2010-01-10: Người dùng u_n đánh giá phim 5

• 2010-01-15: Người dùng u_n đánh giá phim 7

• . . .

• 2010-03-12: Người dùng u_n đánh giá phim 22

Trong trường hợp này, mục tiêu là dự đoán bộ phim mà khách hàng sẽ xếp hạng/lựa chọn kế tiếp dựa trên lịch sử xếp hạng trong quá khứ của khách hàng. Đối với một khách hàng u_n, tạo chuỗi đầu vào $x = (x_1, x_2, \dots, x_T)$ mỗi x_t sẽ biểu diễn cho việc xếp hạng phim. Mỗi x_t là một vectơ mã hóa one-hot, mã hóa phim được khách hàng xếp hạng tại thời điểm t, hoặc phép nhúng biểu diễn bộ phim đó khi sử dụng các phương pháp nhúng. Đưa ra chuỗi đầu vào x_n, T + 1 được dự đoán và biểu diễn cho bộ phim tiếp theo được xếp hạng. Do đó, trong trường hợp này, chúng ta coi vấn đề là phân loại nhiều lớp với chỉ một lớp hợp lệ.

Trong các thử nghiệm này, sẽ sử dụng dữ liệu từ tháng 1 năm 2009 trở đi. Các bộ phim có ít hơn 20 xếp hạng sẽ bị loại bỏ. Cách tạo bộ dữ liệu training và bộ dữ liệu thử nghiệm như sau:

Bộ dữ liệu training: xem xét dữ liệu từ tháng 1 năm 2009 đến tháng 3 năm 2014. Đối với mỗi khách hàng u_n, tạo một mẫu dữ liệu như đã đề cập trong phần 2.3. Điều này có nghĩa là đối với mọi xếp hạng phim, sẽ tạo một mẫu dữ liệu trong đó trình tự đầu vào chứa các xếp hạng trước đó cho đến thời điểm đó và nhãn thực là mã hóa duy nhất hoặc phần nhúng của phim được xếp hạng trong thời điểm thực tế. Hơn nữa chỉ tạo mẫu khi có ít nhất 5 xếp hạng phim trước xếp hạng thực tế và giới hạn số lượng xếp hạng trên mỗi mẫu là 100. Do đó, nếu khách hàng đã xếp hạng hơn 100 xếp hạng thì chỉ xem xét 100 xếp hạng mới nhất. Tập huấn luyện cuối cùng chứa 4053420 mẫu dữ liệu cho tổng số 32901 khách hàng khác nhau.

Bộ dữ liệu thử nghiệm: bộ dữ liệu này sẽ tập trung vào dự đoán xếp hạng của khách hàng từ tháng 4 năm 2014 cho đến tháng 4 năm 2015. Theo đó giai đoạn này là giai đoạn thử nghiệm. Mặc dù việc tối ưu hóa mô hình để tìm hiểu phim nào sẽ được xếp hạng tiếp theo (dự đoán ngắn hạn), tuy nhiên cũng đo lường cuối cùng khách hàng sẽ xếp hạng phim nào (dự đoán dài hạn). Do đó, đối với mỗi khách hàng, sẽ tạo ra các nhãn thực sự $y_n = (y_{n1}, y_{n2}, \dots, y_{nP})$ theo đó y_{nt} là bộ phim thứ

t do khách hàng đánh giá u_n và P là số lượng phim được khách hàng đánh giá trong thời gian thử nghiệm. Để dự đoán các bộ phim được xếp hạng, tiếp tục tạo trình tự $x_n = (x_{n1}, x_{n2}, \dots, x_{nT})$, chứa những bộ phim mà khách hàng đã xem u_n trước thời gian kiểm tra. Sau đó giới hạn phim của chuỗi đầu vào trong 100 phim cuối cùng đánh giá của khách hàng. Dữ liệu kiểm tra cuối cùng chứa 3669 khách hàng khác nhau.

Trong bộ dữ liệu này bao gồm các master data sau:

- Movies: chưa danh sách các bộ phim thuộc nhiều thể loại khác nhau từ năm 2009 đến năm 2015, ví dụ: *"Toy Story (1995)", "Jumanji (1995)", "Tom and Huck (1995)",...*
- Genome_tags: bao gồm thông tin các nhãn được gắn cho các bộ phim, gồm: *"Adventure", "Action", "80s",...*
- Rates: chứa thông tin đánh giá của khách hàng cho từng bộ phim cụ thể, bao gồm nhiều bộ phim được đánh giá bởi nhiều khách hàng theo thang điểm từ 0 đến 5.
- Genome_scores: biểu diễn độ chính xác của việc gán nhãn cho từng bộ phim. Với thang điểm từ 0 đến 1.

1.2.4. *Phương pháp nghiên cứu.*

Nghiên cứu đến dữ liệu có sẵn, quy trình dọn dẹp và phương pháp phân vùng được sử dụng, mô tả các bước cấu hình được thực hiện để xây dựng các mô hình. Phân tích các kết quả thu được và kết luận về việc sử dụng mạng Nơ-ron trong ứng dụng phân tích hành vi của khách hàng, đề xuất các hướng nghiên cứu tiếp theo.

CHƯƠNG 2. MẠNG RNN VÀ KỸ THUẬT PHÂN TÍCH

Trong chương này, luận văn sẽ trình bày nghiên cứu về các lý thuyết phục vụ cho luận văn, như mạng RNN, các phép nhúng tuyến tính và phi tuyến tính, các cơ chế chú ý tuyến tính và phi tuyến tình nhằng phân tích và đưa ra các dự đoán về hành vi của khách hàng.

2.1. Mạng Nơron và các vấn đề cơ bản.

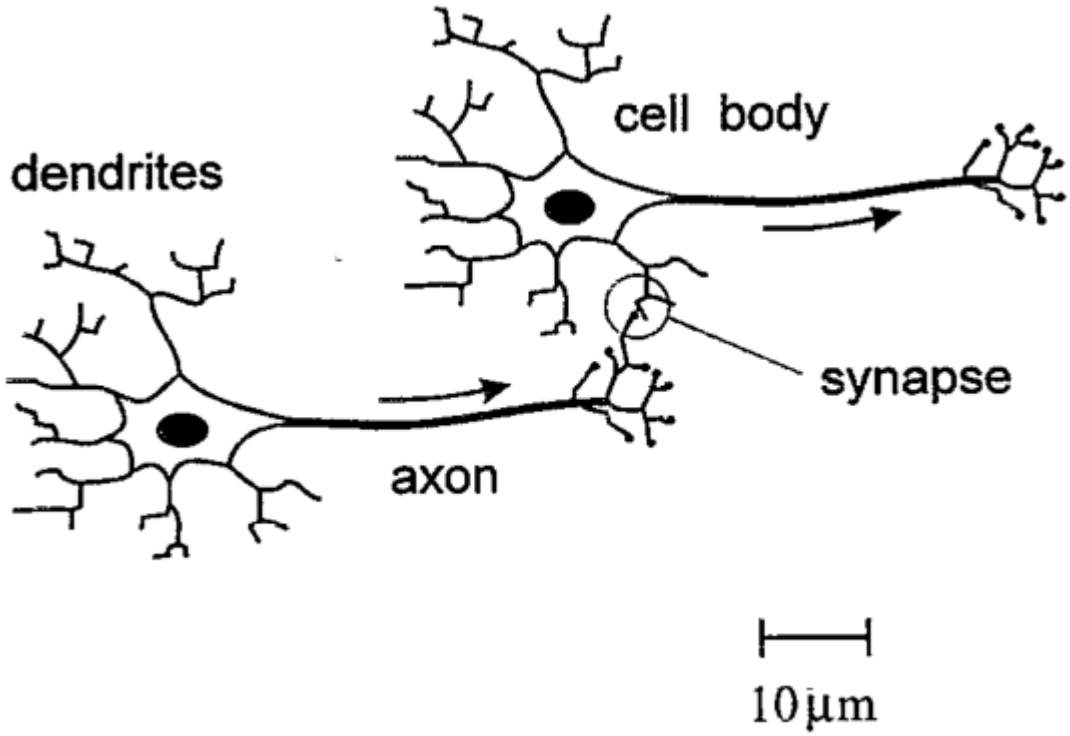
2.1.1. *Tổng quan về mạng lưới thần kinh*

Cách tiếp cận thông thường đối với tính toán dựa trên một tập hợp các hướng dẫn được lập trình rõ ràng và ngày tháng từ công việc của Babbage, Turing và von Neumann. Mạng Nơron đại diện cho một mô hình tính toán thay thế trong đó giải pháp cho một vấn đề được học từ một tập hợp các ví dụ. Nguồn cảm hứng cho mạng lưới thần kinh ban đầu xuất phát từ các nghiên cứu về cơ chế xử lý thông tin trong hệ thần kinh sinh học, đặc biệt là não người. Thật vậy, phần lớn các nghiên cứu hiện tại về các thuật toán mạng Nơron tập trung vào việc hiểu sâu hơn về xử lý thông tin trong các hệ thống sinh học. Tuy nhiên, các khái niệm cơ bản cũng có thể được hiểu từ một cách tiếp cận trừu tượng thuần túy để xử lý thông tin. Đối với tính đầy đủ, chúng ta sẽ có một cái nhìn tổng quan ngắn gọn về mạng thần kinh sinh học trong cuối chương này. Tuy nhiên, trọng tâm trong luận văn này sẽ chủ yếu tập trung vào các mạng nhân tạo cho các ứng dụng thực tế.

Một mạng Nơron truyền thẳng có thể được coi là một hàm toán học phi tuyến tính biến đổi một tập hợp đầu vào các biến thành một tập hợp các biến đầu ra. Dạng chính xác của sự biến đổi được điều chỉnh bởi một tập hợp các tham số được gọi là trọng số mà giá trị có thể được xác định trên cơ sở một tập hợp ví dụ về ánh xạ bắt buộc. Quá trình xác định các giá trị tham số này thường được gọi là học hỏi hoặc đào tạo, và có thể là một công việc chuyên sâu về tính toán. Khi các trọng số đã được cố định, dữ liệu mới có thể được xử lý bởi mạng lưới một cách rất nhanh chóng. Chúng ta sẽ thấy thuận tiện ở một số điểm trong phần đánh giá của luận văn này sau khi rút ra được sự tương tự giữa mạng Nơron nhân tạo và kỹ thuật tiêu chuẩn của việc điều chỉnh đường cong bằng cách sử dụng các hàm đa thức. Một đa

thức có thể được coi là một ánh xạ từ một đầu vào duy nhất biến thành một biến đầu ra duy nhất. Các hệ số trong đa thức tương tự như trọng số trong mạng Nơron, và việc xác định các hệ số này (bằng cách giảm thiểu tổng bình phương lỗi) tương ứng với quá trình mạng đào tạo.

Ngoài việc cung cấp tốc độ xử lý cao, mạng Nơron có khả năng quan trọng là học hỏi từ một giải pháp chung cho một vấn đề từ một tập hợp các ví dụ cụ thể.



Hình 2.1. Sơ đồ minh họa của hai tế bào thần kinh sinh học.

Hành động của các đuôi gai làm đầu vào và khi một tế bào thần kinh kích hoạt một điện thế hoạt động sẽ lan truyền sợi trục theo hướng được chỉ ra bởi mũi tên. Sự tương tác giữa các tế bào thần kinh diễn

2.1.1.1. *Mạng lưới thần kinh sinh học*

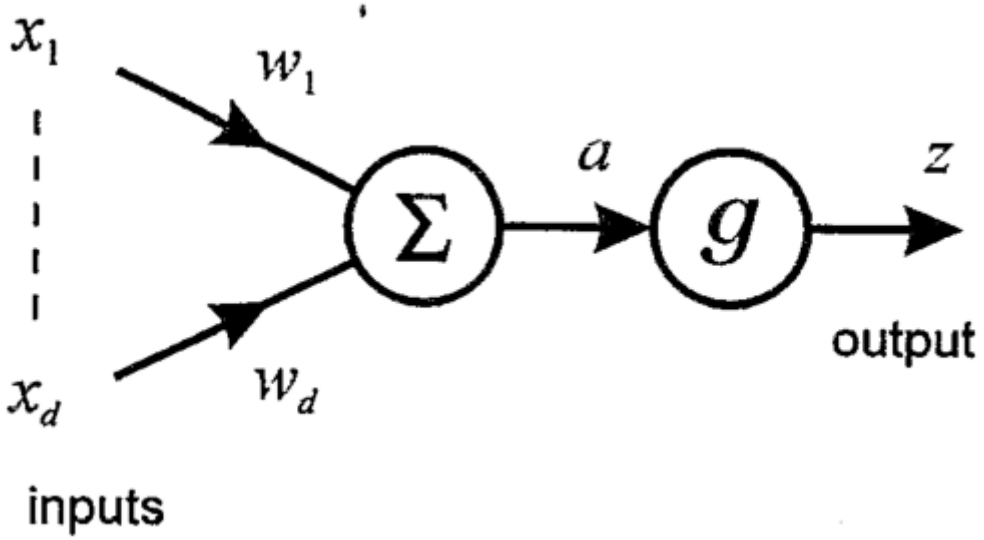
Bộ não con người là cấu trúc phức tạp nhất được biết đến, và hiểu cách thức hoạt động của nó là một trong những những thách thức khó khăn và thú vị mà khoa học phải đối mặt. Mạng thần kinh sinh học cung cấp động lực đằng sau rất nhiều nghiên cứu về các mô hình mạng nhân tạo, nhằm bổ sung cho mong muốn xây dựng khả năng nhận dạng mẫu tốt hơn và hệ thống xử lý thông tin.

Bộ não con người chứa khoảng 10^{11} tế bào hoạt động điện được gọi là tế bào thần kinh. Chúng tồn tại trong một loạt các các hình thức khác nhau, mặc dù hầu hết đều có các tính năng chung được chỉ ra trong Hình 2.1. Cây phân nhánh của cây đuôi gai cung cấp một tập hợp các đầu vào cho nơron, trong khi sợi trục hoạt động như một đầu ra. Giao tiếp giữa các tế bào thần kinh diễn ra tại các điểm nối được gọi là khớp thần kinh. Mỗi tế bào thần kinh thường tạo ra kết nối đến hàng nghìn Nơron khác, do đó, tổng số số lượng khớp thần kinh trong não vượt quá 10^4. Mặc dù mỗi Nơron là một hệ thống xử lý thông tin tương đối chậm (hoạt động trên thang thời gian hiệu dụng khoảng 1 ms). Việc xử lý thông tin khổng lồ song song tại nhiều khớp thần kinh đồng thời dẫn đến một sức mạnh xử lý hiệu quả vượt xa nhiều so với siêu máy tính ngày nay. Nó cũng dẫn đến mức độ chịu lỗi cao, với nhiều tế bào thần kinh chết mỗi ngày với ít ảnh hưởng xấu đến hiệu suất.

Nhiều tế bào thần kinh hoạt động theo cách tất cả hoặc không có gì, và khi các tế bào này "kích hoạt", chúng gửi một xung điện (được gọi là một hành động tiềm năng) lan truyền từ cơ thể tế bào dọc theo sợi trục. Khi tín hiệu này đến khớp thần kinh, nó sẽ kích hoạt giải phóng các chất dẫn truyền thần kinh hóa học đi qua khớp thần kinh điểm nối với Nơron tiếp theo. Tùy thuộc vào loại khớp thần kinh, điều này có thể tăng (khớp thần kinh kích thích) hoặc giảm (khớp thần kinh ức chế) xác suất của tế bào thần kinh tiếp theo sự khai hỏa. Mỗi khớp thần kinh có một sức mạnh (hoặc trọng lượng) liên quan xác định độ lớn của tác động của một xung lực trên nơron sau khớp thần kinh. Do đó, mỗi tế bào thần kinh sẽ tính toán tổng trọng số của các đầu vào từ các tế bào thần kinh khác, và nếu tổng số kích thích vượt quá ngưỡng nào đó, tế bào thần kinh sẽ kích hoạt. Như chúng ta sẽ thấy ở phần sau, các mạng nơron như vậy có khả năng xử lý thông tin rất tổng quát.

Một đặc tính quan trọng của cả hệ thống thần kinh thực và hệ thần kinh nhân tạo là khả năng sửa đổi phản ứng của chúng do tiếp xúc với các tín hiệu bên ngoài. Điều này thường được gọi là học tập, và chủ yếu xảy ra thông qua những thay đổi về sức mạnh của các khớp thần kinh. Bức tranh về các hệ thống thần kinh sinh học được đơn giản hóa ở trên cung cấp một điểm khởi đầu thuận tiện cho cuộc thảo luận về

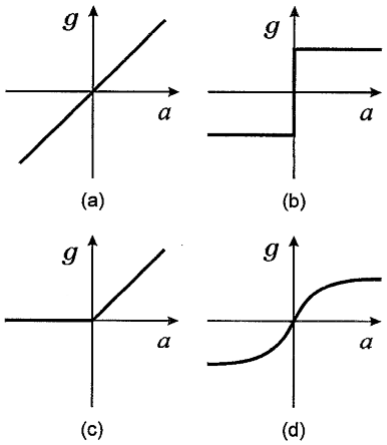
các mô hình mạng nhân tạo.



Hình 2.2. Mô hình McCulloch-Pitts của một tế bào thần kinh đơn lẻ.

2.1.1.2. *Mạng lưới thần kinh nhân tạo*

Một mô hình toán học đơn giản của một tế bào thần kinh đơn lẻ được giới thiệu trong một bài báo của McCulloch và Pitts trong 1943, và có hình dạng như trong Hình 2.3. Nó có thể được mô tả lại như một hàm phi tuyến tính biến đổi một tập hợp biến đầu vào x_i, (i = 1, ..., d) vào biến đầu ra z. Lưu ý rằng từ bây giờ chúng ta sẽ tham khảo một mô hình nhân tạo của một tế bào thần kinh như một đơn vị xử lý, hoặc đơn giản hơn, để phân biệt nó với đối tác sinh học của nó.



Hình 2.3. Lựa chọn các chức năng kích hoạt điển hình: (a) tuyến tính, (b) ngưỡng, (c) ngưỡng tuyến tính, (d) tuyến tính.

Trong mô hình McCulloch-Pitts, tín hiệu x_i ở đầu vào i là lần đầu tiên được nhân với một tham số w_iđược gọi là trọng số (tương tự như sức mạnh của khớp thần kinh trong một mạng lưới sinh học) và sau đó được thêm vào tất cả các tín hiệu đầu vào có trọng số khác, để cung cấp tổng đầu vào cho đơn vị của biểu mẫu

$$a = \sum_{i=1}^{d} w_i x_i + w_0 \tag{2.1}$$

Trong đó tham số bù w_0 được gọi là sai lệch (tương ứng với ngưỡng mệt mỏi trong một tế bào thần kinh sinh học). Về mặt hình thức, độ sai lệch có thể được coi là một trường hợp đặc biệt của trọng số từ một đầu vào bổ sung có giá trị x_0 được đặt vĩnh viễn thành + 1. Như vậy chúng ta có thể viết Eq. (1) dưới dạng

$$\mathrm{a} = \sum_{\mathrm{i}=0}^{\mathrm{d}} \mathrm{w_i x_i} \tag{2.2}$$

Trong đó $x_0 = 1$. Lưu ý rằng trọng số(và các sai lệch) có thể là một trong hai dấu hiệu, tương ứng với các khớp thần kinh hưng phấn hoặc ức chế. Đầu ra z của đơn vị (có thể được coi là tương tự như tốc độ mệt mỏi trung bình của một tế bào thần kinh) sau đó được đưa ra bằng cách hoạt động trên a với hàm kích hoạt phi tuyến tính g () sao cho:

$$\mathrm{z = g(a)} \tag{2.3}$$

Một số hình thức có thể cho hàm g () được hiển thị trong Hình. 2.3. Mô hình McCulloch-Pitts ban đầu sử dụng ngưỡng chức năng được hiển thị trong Hình 2.3 (b). Hầu hết các mạng quan tâm thiết thực đều sử dụng sigmoidal (có nghĩa là kích hoạt hình chữ S! Các chức năng của loại được hiển thị trong Hình 2.3 (d).

Như chúng ta sẽ thấy, mô hình Nơron đơn giản này hình thành yếu tố toán học cơ bản trong nhiều mô hình mạng Nơron nhân tạo. Bằng cách liên kết với nhiều phần tử xử lý đơn giản như vậy, có thể tạo ra một lớp rất chung chung của ánh xạ phi tuyến tính, có thể được áp dụng cho một phạm vi rộng của các vấn đề thực tế. Việc điều chỉnh các giá trị trọng số, theo một thuật toán huấn luyện thích hợp, có thể cho phép các mạng học hỏi theo phản ứng với dữ liệu bên ngoài. Mặc dù chúng ta đã

giới thiệu mô hình toán học này của tế bào thần kinh như một đại diện cho hoạt động của các tế bào thần kinh sinh học, chính xác là những ý tưởng tương tự cũng nảy sinh khi chúng ta xem xét các cách tiếp cận tối ưu để giải quyết các vấn đề trong nhận dạng mẫu thống kê.

2.2. Phương pháp nghiên cứu.

2.2.1. *Mạng RNN cơ sở*

Đầu tiên bắt đầu với mô hình RNN không có thuật toán nhúng và cơ chế chú ý, được gọi là RNN-baseline.

Mỗi phần tử x_t của chuỗi đầu vào $x = (x_1, x_2, \dots, x_T)$ là một vectơ điểm nóng hoặc mã hóa một hoặc nhiều điểm nóng. Mỗi phần tử được đưa vào một khối LSTM, tạo ra một vectơ trạng thái ẩn h_t:

$$h_t = LSTM(x_t, h_{t-1}) \tag{2.4}$$

Sau khi xử lý toàn bộ dãy và thu được vectơ trạng thái ẩn cuối cùng h_T, chúng ta thu được các xác suất dự đoán như sau:

$$\hat{y}_{T+1} = g(W_{out} h_T + b_{out-1}) \tag{2.5}$$

Khi giải quyết một vấn đề nhiều lớp với một lớp định nghĩa duy nhất cho mỗi mẫu, g là một hàm softmax. Sau đó, việc tối ưu hóa trọng số của mô hình W_f, b_f, W_i, b_i, W_o , b_o trong khối LSTM và W_{out} trong lớp cuối cùng được thực hiện bằng cách giảm thiểu entropy chéo của mục chính xác của mẫu:

$$L = \frac{1}{N}(\sum_{n=1}^{N} y_{T+1} \log(\hat{y}_{T+1}) \tag{2.6}$$

Việc giảm độ dốc theo lô nhỏ được sử dụng, đối với mỗi lần lặp lại giảm biên độ N đại diện cho kích thước một batch, tương ứng với số lượng khách hàng trong minibatch.

Bảng 2.1. Các phương pháp khác nhau sử dụng phép nhúng.

RNN-Emb-Word2vec	Phép nhúng kết hợp với Word2vec.
RNN-Emb-Jointly-Lin	Phép nhúng kết hợp phân loại tuyến tính.
RNN-Emb-Jointly-Nonlin	Phép nhúng kết hợp phân loại phi tuyến tính.
RNN-Emb-Word2vec-Finetune	Phép nhúng riêng biệt và được điều chỉnh vi cấp chung
RNN-Emb-Output	Phép nhúng được đào tạo riêng biệt

Khi giải quyết vấn đề phân loại nhiều nhãn với nhiều lớp hợp lệ trên mỗi mẫu, g là hàm sigmoid và được tối ưu hóa trọng số bằng cách giảm thiểu entropy chéo, tính trung bình sự mất mát của tất cả các lớp:

$$L = \frac{1}{N}(\sum_{n=1}^{N} y_{T+1}\log(\hat{y}_{T+1}) + (1 - y_{T+1})\log(1 - \hat{y}_{T+1}) \quad (2.7)$$

2.2.2. *Các phương pháp nhúng*

2.2.2.1. *Phép nhúng được đào tạo riêng với Word2vec*

Như chúng ta đã thấy, phép nhúng đã được áp dụng thành công trong nhiều tác vụ NLP. Cũng như trong công việc xử lý chuỗi các đầu mục, có thể áp dụng cùng một cách tiếp cận để tạo ra các phép nhúng các item.

Đầu tiên, nếu có một số lượng lớn các item, kích thước của mỗi mặt hàng x_t khi sử dụng việc mã hóa vectơ one-hot có thể rất lớn. Thứ tự của đầu vào được kết nối trực tiếp với khối LSTM. Kết quả là số lượng trọng số trong các cổng W_i, W_f, và W_o tăng trưởng cùng với số lượng item. Do đó, mô hình có nhiều tham số để học hỏi và yêu cầu nhiều dữ liệu hơn để được huấn luyện thành công. Hơn nữa, ma trận phép nhân khổng lồ có thể làm cho việc huấn luyện và dự đoán bị chậm lại. Ánh xạ item thành một vectơ giá trị thực có nhiều chiều hơn thay vì một vectơ mã hóa one-hot khổng lồ có thể giúp giải quyết vấn đề này.

Lý do thứ hai là mã hóa one-hot không cung cấp bất kỳ thông tin tương tự hoặc ngữ nghĩa nào về đầu vào. Bằng cách sử dụng các phương pháp hiệu quả, chúng ta có thể trình bày các phép nhúng item chứa thông tin ngữ nghĩa về mặt hàng, điều này có thể giúp tăng hiệu suất dự đoán vì chúng ta đang bao gồm thông tin bổ sung

về mặt hàng.

Một lợi thế khác nữa của việc sử dụng các phép nhúng là nếu số lượng các mục tăng lên thì kích thước của biểu diễn phép nhúng không thay đổi.

Để áp dụng phương pháp nhúng cho các mục, việc thay thế các từ(câu) được biểu diễn trong word2vec bằng chuỗi các mục mà một khách hàng cụ thể đã sử dụng. Phần còn lại của phương pháp được áp dụng cùng phương pháp Skip-gram và lấy mẫu thụ động.

Trong [5], thông tin không gian mẫu được loại bỏ bằng cách coi kích thước cửa sổ là cùng một kích thước của toàn bộ chuỗi mục khách hàng và đưa ra cùng một trọng số cho tất cả các mục.

Các phép nhúng đầu mục sử dụng Skip-gram với phương pháp lấy mẫu thụ động được tạo như sau:

Cho một chuỗi các mục $(i_i)_{i=0}^{T}$ mà khách hàng đã sử dụng, mục tiêu là tối đa hóa:

$$\frac{1}{K} = (\sum_{i=1}^{K} \sum_{-c \le j \le c, j \ne 0} \log p(i_{i+j}|i_i) \tag{2.8}$$

trong đó c là kích thước cửa sổ ngữ cảnh và số $p(i_{i+j}|i_i)$ được định nghĩa là:

$$p(i_j|i_i) = (\sigma(u_i^T v_j) \prod_{k=1}^{N} \sigma(-u_i^T v_k) \tag{2.9}$$

Ở đây $\sigma(x) = \frac{1}{1} + \exp(-x)$ và N là số lượng các mẫu thụ động để rút ra mỗi mẫu chủ động. Một item thụ động i_i được lấy mẫu từ phân phối unigram được nâng lên lũy thừa 3/4. Các vectơ $u_i \in U(\subset R^m)$ và $v_i \in V(\subset R^m)$ tương ứng với mục tiêu và đại diện ngữ cảnh cho mặt hàng i_i, tương ứng, và m là kích thước của phép nhúng.

Để chống lại sự mất cân bằng giữa các item hiếm và các item thường xuyên, mỗi item i_i trong chuỗi đầu vào bị loại bỏ với xác suất theo công thức:

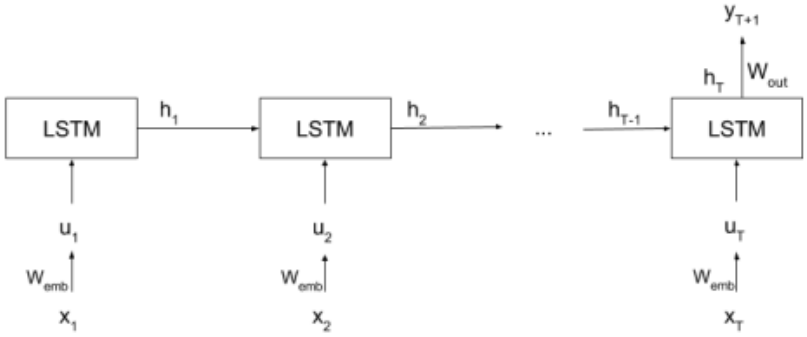
$$p(discard|i_i) = 1 - \sqrt{\frac{t}{f(i_i)}} \quad (2.10)$$

Ở đây $f(i_i)$ là tần suất của item i_i và t là một ngưỡng đã được chọn. Một lần tối ưu hóa bằng cách tối đa hóa 4.5, u_i được sử dụng để nhúng item i_i.

Khi có được các mục nhúng, mô hình có thể được mô tả là mô hình RNN-Baseline, với sự khác biệt của mỗi phần tử x_i của chuỗi đầu vào x = (x_1, x_2, ..., x_T) đại diện cho phép nhúng mục tương ứng u_i. Giá trị M được biểu thị là phần thứ nguyên của các phép nhúng và K là tổng số item. Như M << K, thì phần thứ nguyên của ma trận trọng số W_i, W_f, W_o sẽ thấp hơn nhiều so với mô hình trước đó và thu được một mô hình với ít tham số hơn để học hỏi. Hơn nữa, nhằm tránh việc tính toán phép nhân ma trận có kích thước lớn, mô hình RNN-Emb-Word2vec sẽ được tham chiếu và sử dụng.

2.2.2.2. *Phép nhúng được tìm hiểu với mô hình phân loại*

Mặc dù tìm hiểu phép nhúng bằng một phương pháp riêng biệt có thể tạo ra các biểu diễn vectơ chất lượng cao, nhưng chúng có thể không được tối ưu hóa cho nhiệm vụ cụ thể cần giải quyết. Vì lý do này, một cách thức được đề xuất tìm hiểu các biểu diễn phép nhúng cùng với mô hình phân loại. Hình 2.4 mô tả mô hình. Như trong mô hình RNN-Baseline mô hình hóa từng phần tử x_i đại diện cho một vectơ mã hóa one-hot hoặc multi-hot. Trong trường hợp này, phần tử được ánh xạ tới một phép nhúng theo trọng số ma trận nhúng W_{emb}.



Hình 2.4. Mô hình RNN-Emb-Jointly-Lin and RNN-Emb-Jointly-Nonlin.

Ở đây chúng ta xem xét hai lựa chọn. Học phép nhúng theo hàm tuyến tính của đầu vào:

$$u_t = W_{emb} x_t \tag{2.11}$$

Và học phép nhúng bằng một hàm phi tuyến tính của đầu vào:

$$u_t = \tanh(W_{emb} x_t + b_{emb}) \tag{2.12}$$

Kết quả của phép nhúng u_t được sử dụng làm đầu vào của khối LSTM. Phần còn lại của mô hình tương tự như RNN-Baseline. Như RNN-Emb-Word2vec ma trận trọng số W_i, W_f, W_o có kích thước thấp hơn nhiều so với RNN-Baseline, nhưng có một ma trận trọng số mới W_{emb} giảm kích thước của M × K để tìm hiểu. Được tham chiếu tới mô hình RNN-Emb-Jointly-Lin khi sử dụng nhúng phi tuyến tính.

2.2.2.3. *Phép nhúng được tìm hiểu học riêng biệt và sau đó được điều chỉnh chung.*

Việc học phép nhúng chất lượng cao cùng với mô hình phân loại từ đầu có thể là một nhiệm vụ phức tạp. Trong phương pháp này, những gì tốt nhất của các mô hình trước đó sẽ được nghiên cứu và sử dụng. Kỹ thuật này đã được áp dụng thành công trong[9] trong bối cảnh y tế. Mô hình có thể được biểu diễn dưới dạng RNN-Emb-Jointly-Lin với mô hình phép nhúng tuyến tính thể hiện trong hình 2.4. Điểm khác biệt là ma trận nhúng W_{emb} đã được khởi tạo với các phép nhúng đã được đào tạo trước đó thông qua phương pháp Skip-gram.

2.2.2.4. *Các phép nhúng dự đoán. Các phép nhúng được học riêng*

Trong các mô hình trước, chỉ tập chung sử dụng các phép nhúng trước khối LSTM. Trong[7] một mô hình mới đã được giới thiệu trong đó dự đoán cuối cùng của mô hình là một phép nhúng. Trong các mô hình trước, ma trận trọng số W_{out} được kết nối từ trạng thái ẩn của khối LSTM đến đầu ra của mạng, trong đó xác suất của từng giá trị là khác nhau. Nếu K là số lượng các items và H là số lượng các node trong trạng thái ẩn, sau đó là số lượng tham số của W_{out} là K × H, phát triển theo số lượng items.

Như trong[7], một mô hình dự đoán phép nhúng items được tiêu thụ tiếp theo

được đề xuất, thay vì dự đoán xác suất tiêu thụ từng mặt. Nếu M là kích thước của các phép nhúng, thì phương pháp này làm giảm số lượng tham số của W_{out} điều đó làm giảm số lượng tham số xuống M × H, với M << K. Bên cạnh việc giảm số lượng tham số, tránh tính toán một ma trận nhân lớn và xử lý softmax, do đó thời gian dự đoán có thể được giảm xuống như trong[7].

Mô hình được đào tạo bằng cách giảm thiểu sự mất cosin giữa phép nhúng với đầu ra thực tế và phép nhúng được dự đoán. Một trong những yêu cầu để đạt được độ chính xác tốt là sử dụng các phép nhúng có hiệu quả tốt. Trong [7], mô hình được báo cáo là hoạt động kém hơn so với các mô hình không có sử dụng phép dự đoán nhúng về độ chính xác của dự đoán.

Mô hình có thể được biểu diễn dưới dạng RNN-Baseline trong hình 2.1. Trong trường hợp này, mỗi phần tử x_t của chuỗi đầu vào đại diện cho phép nhúng của một item, được kết nối trực tiếp với khối LSTM theo cùng một cách với mô hình RNN-Baseline. Mỗi trạng thái ẩn h_t thu được thông qua các phương trình LSTM. Cuối cùng, phép dự đoán nhúng được tính như sau:

$$\hat{y}_{T+1} = W_{out} h_T + b_{out}) \tag{2.13}$$

Kích thước của y_{T+1} là kích thước nhúng kích thước M trong trường hợp này.

Hàm mất mát được tối thiểu hóa là tổn thất cosin giữa lần nhúng dự đoán $\hat{y}_{T+1}$ và nhúng thực sự y_{T+1}:

$$L = \frac{1}{N}\sum_{n=1}^{N}(1 - \frac{\hat{y}_{T+1} \cdot y_{T+1}}{||\hat{y}_{T+1}||2 \cdot ||y_{T+1}||2}) \tag{2.14}$$

Mô hình này được gọi là RNN-Emb-Output.

2.2.3. *Cơ chế chú ý*

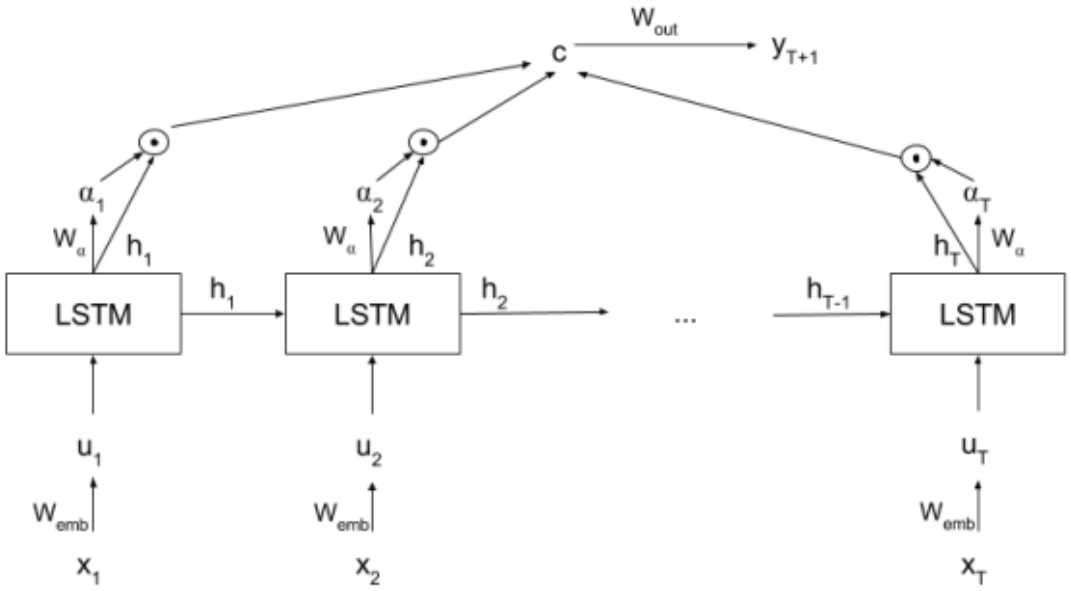
Như đã đề cập trước đây, các cơ chế chú ý có thể được sử dụng để giải thích các dự đoán của mô hình, cho thấy tầm quan trọng của các yếu tố quá khứ khác nhau của quy trình đầu vào khi đưa ra mỗi dự đoán. Hơn nữa, trong các mô hình trước, các dự đoán đã đưa ra chỉ sử dụng trạng thái ẩn cuối cùng, nghĩa là mô hình phải học cách chuyển tất cả các thông tin cần thiết qua tất cả các chuỗi để đưa ra dự

đoán, điều này có thể là rào cản với các chuỗi dài. Với cơ chế chú ý, một vectơ ngữ cảnh được tạo ra bằng cách tập trung vào các phần tử khác nhau của chuỗi đầu vào, điều này tránh được việt bắt buộc phải mã hóa tất cả thông tin ở trạng thái ẩn cuối cùng.

Đối với các mô hình chú ý, mô hình cơ sở RNN-Emb-Word2vec-Finetune được sử dụng như một cơ sở chính, chứa một lớp các phép nhúng giữa đầu vào và khối LSTM. Điều này sẽ được trình bày trong chương 3 nhằm chỉ ra tất cả các phương pháp nhúng đều đạt được hiệu suất tương tự khi được đào tạo. Tuy nhiên, RNN-Emb-Word2vec-Finetune có tính tập trung nhanh hơn. Vì lý do này, nó được sử dụng làm cơ sở cho các mô hình chú ý. Bảng 2.2 tóm tắt các cơ chế chú ý khác nhau. Trong các phần tiếp theo, các phương pháp này sẽ giới thiệu chi tiết.

Bảng 2.2. Các phương thức khác nhau sử dụng phương phức chú ý.

RNN-Att-HS-Lin	Cơ chế chú ý đến các trạng thái RNN ẩn với trọng số chú ý tuyến tính
RNN-Att-HS-Nonlin	Cơ chế chú ý đến các trạng thái RNN ẩn với trọng số chú ý phi tuyến tính
RNN-Att-Emb-Lin	Cơ chế chú ý trỏ đến các phép nhúng có trọng số tuyến tính
RNN-Att-Emb-Nonlin	Cơ chế chú ý trỏ đến các phép nhúng có trọng số phi tuyến tính



Hình 2.5. Mô hình RNN-Att-HS-Lin and RNN-Att-HS-Nonlin. Khối LSTM với cơ chế chú ý đến trạng thái ẩn.

2.2.3.1. *Cơ chế chú ý đến các trạng thái ẩn RNN*

Gần đây, sự chú ý đến trạng thái ẩn đã được áp dụng thành công trong một số tác phẩm[3, 31, 44]. Ý tưởng là tạo một vectơ định nghĩa văn bản bằng cách tập trung chú ý vào các trạng thái ẩn khác nhau được tạo ra trong khi xử lý chuỗi đầu vào. Mô hình được thể hiện trong hình 2.5. Sự khác biệt RNN-Emb-Word2vec-Finetune là chúng ta tạo ra một trọng số $\propto_t$ cho mọi trạng thái ẩn h_t. Các trọng số này thể hiện mức độ tập trung của mô hình ở mọi trạng thái ẩn để tạo vectơ định nghĩa văn bản c, được sử dụng để dự đoán mặt hàng nào sẽ được tiêu thụ tiếp theo.

Mỗi trọng số $\propto_t$ được tạo như sau. Đầu tiên, một năng lượng e_t được tính toán, trong đó một hàm tuyến tính của trạng thái ẩn được xem xét:

$$e_t = W_{\propto} h_t + b_{\propto} \tag{2.15}$$

và hàm phi tuyến tính bằng cách sử dụng tanh:

$$e_t = \tanh(W_{\propto} h_t + b_{\propto}) \tag{2.16}$$

Sau đó, tiến hành tính toán trọng số $\propto_t$ bằng cách sử dụng hàm softmax với tất cả các phần tử e_t:

$$\propto_t = \frac{exp\ e_t}{\sum_i^T e_i} \tag{2.17}$$

Để tạo vectơ định nghĩa văn bản c, trọng số $\propto_t$ được sử dụng để tính giá trị trọng số trung bình của các trạng thái ẩn:

$$c = \sum_{t=1}^{T} \propto_t h_t \tag{2.18}$$

Cuối cùng, vectơ định nghĩa văn bản được sử dụng để dự đoán nhãn cho chuỗi đầu vào:

$$\hat{y}_{T+1} = g(W_{out} c + b_{out}) \tag{2.19}$$

Như trong các mô hình trước, g là hàm sigmoid cho phân loại đa nhãn với nhiều lớp hợp lệ trên mỗi mẫu và hàm softmax cho phân loại nhiều lớp chỉ có một lớp hợp lệ. Một hàm mất mát được sử dụng chung cho mô hình RNN-Baseline. Mô hình này

được biểu thị là RNN-Att-HS-Lin khi sử dụng hàm tuyến tính để tính toán e_t và RNN-Att-HS-Nonlin khi sử dụng một hàm phi tuyến tính.

2.2.3.2. *Cơ chế chú ý kết hợp với các phép nhúng*

Cơ chế chú ý đến trạng thái ẩn là hình thức phổ biến nhất trong văn học. Gần đây, [2] đã đề xuất một phương pháp chú ý trong đó trọng số chú ý được sử dụng với các phép nhúng. Dựa vào nghiên cứu đó, phương pháp được trình bày trong hình 2.6 được đề xuất. Sự khác biệt giữa mô hình RNN-Att-Emb-Lin và RNN-Att-Emb-Nonlin là vectơ định nghĩa văn bản được tạo ra bằng cách tập trung sự chú ý kết hợp với các phép nhúng thay vì các trạng thái ẩn. Mô hình này đơn giản hơn mô hình được sử dụng trong [2], trong đó hai mạng RNN được đào tạo để tạo ra 2 bộ trọng số chú ý khác nhau, một bộ cho mức độ truy cập và một bộ khác cho mức độ thay đổi.

Trọng số chú ý $\propto_t$ được tính theo cùng một cách so với các mô hình RNN-Att-HS-Lin và RNN-Att-HS-Nonlin. Trong trường hợp này, vectơ định nghĩa văn bản c được tạo ra bằng cách tính toán trọng số trung bình của các phép nhúng khác nhau dựa trên dữ liệu đầu vào:

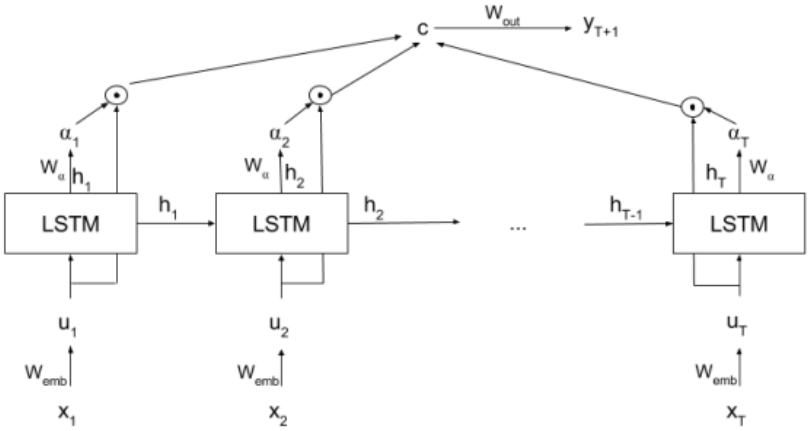
$$c = \sum_{t=1}^{T} \propto_t u_t \tag{2.20}$$

Phần còn lại của mô hình giống với RNN-Att-HS-Lin và RNN-Att-HS-Nonlin. Sẽ tham chiếu từ mô hình như RNN-Att-Emb-Lin khi sử dụng một hàm tuyến tính để kết hợp tính toán e_t và RNN-Att-Emb-Nonlin khi sử dụng một hàm phi tuyến tính.

Tập trung vào các phép nhúng cho phép tạo ra vectơ định nghĩa văn bản bằng cách tập trung hoàn toàn vào các phần tử khác nhau của chuỗi, vì mỗi phép nhúng chỉ chứa thông tin của chính nó . Ngược lại, các trạng thái ẩn khác nhau có thể chứa thông tin từ các đầu vào trước đó. Do đó, việc chú ý kết hợp với các phép nhúng có thể là một phương pháp tốt hơn để biết mạng RNN xem xét đến những mức độ nào đối với từng đầu vào trước đó khi đưa ra dự đoán.

2.3. Xử lý dữ liệu tuần tự

Khi xử lý dữ liệu tuần tự, Nhiều cách thức khác nhau có thể được sử dụng để tạo ra các mẫu dữ liệu. Xét dãy gồm 4 phần tử $s = (s_1, s_2, s_3, s_4)$ xem xét ba lựa chọn tiếp theo:



Hình 2.6. Mô hình RNN-Att-Emb-Lin và RNN-Att-Emb-Nonlin.

- Tạo một mẫu trên mỗi chuỗi hoàn chỉnh, xem xét việc mất mát của yếu tố cuối cùng của chuỗi. Nó tạo ra mẫu $x = (s_1, s_2, s_3)$ có nhãn $y = s_4$. Trong trường hợp này, mô hình có thể học hỏi được điều đó sau chuỗi (s_1, s_2, s_3) đến từ phần tử s_4, nhưng nó không học được về các yếu tố trung gian.
- Tạo một mẫu trên mỗi chuỗi hoàn chỉnh, nhưng xem xét phần trung gian. Nó tạo ra mẫu $x = (s_1, s_2, s_3)$ có nhãn $y = (s_2, s_3, s_4)$. Trong trường hợp này, mô hình có thể học được với các phần tử trung gian bằng cách sao chép lại lỗi ở mỗi bước theo thời gian.
- Tạo một mẫu cho mỗi tiền tố trước của chuỗi. Phương pháp này đã được sử dụng trong [7] và được gọi là tăng cường dữ liệu. Nó bao gồm việc tạo một mẫu khác nhau cho từng tiền tố trước của chuỗi. Trong ví dụ này, tạo các mẫu tiếp theo:

  $x = (s_1), y = (s_2)$

  $x = (s_1, s_2), y = (s_3)$

  $x = (s_1, s_2, s_3), y = (s_4)$

  Như tùy chọn trước, điều này cho phép mô hình học hỏi từ các phần tử trung gian.

Sau khi đã thử ba tùy chọn trong các thử nghiệm ban đầu. Vì tùy chọn thứ ba tạo ra kết quả tốt nhất, nên phương pháp tiền xử lý này sẽ được sử dụng cho các thử nghiệm tiếp theo, bao gồm các thử nghiệm được báo cáo trong chương 3.

CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ.

Trong phần này, tập trung vào việc trình bày các tập dữ liệu được sử dụng trong việc phân tích, thiết lập công cụ và các thử nghiệm được thực hiện và các chỉ số đánh giá để đo lường hiệu suất của các mô hình.

3.1. Bộ dữ liệu

Tập dữ liệu được sử dụng là tập dữ liệu danh sách các bộ phim được bán ra và được đánh giá bởi khách hàng từ năm 2008 đến năm 2015[1]. Tập dữ liệu bao gồm lịch sử xếp hạng các bộ phim do khách hàng khác nhau đánh giá. Bộ dữ liệu xếp hạng chứa một giá trị là mốc thời gian, cho biết thứ tự mà khách hàng xếp hạng các bộ phim. Trong tập dữ liệu này, các bộ phim được xếp hạng theo điểm số từ 1 đến 5. Tuy nhiên, trong luận văn này chỉ sử dụng các dữ liệu đã khách hàng đã xếp hạng một bộ phim chứ không phải điểm số. Ví dụ về lịch sử tương tác của khách hàng có thể được mô tả như sau:

• 2010-01-10: Người dùng u_n đánh giá phim 5

• 2010-01-15: Người dùng u_n đánh giá phim 7

• . . .

• 2010-03-12: Người dùng u_n đánh giá phim 22

Trong trường hợp này, mục tiêu là dự đoán bộ phim mà khách hàng sẽ xếp hạng/lựa chọn kế tiếp dựa trên lịch sử xếp hạng trong quá khứ của khách hàng. Đối với một khách hàng u_n, tạo chuỗi đầu vào $x = (x_1, x_2, \dots, x_T)$ mỗi x_t sẽ biểu diễn cho việc xếp hạng phim. Mỗi x_t là một vectơ mã hóa one-hot, mã hóa phim được khách hàng xếp hạng tại thời điểm t, hoặc phép nhúng biểu diễn bộ phim đó khi sử dụng các phương pháp nhúng. Đưa ra chuỗi đầu vào x_n, T + 1 được dự đoán và biểu diễn cho bộ phim tiếp theo được xếp hạng. Do đó, trong trường hợp này, chúng ta coi vấn đề là phân loại nhiều lớp với chỉ một lớp hợp lệ.

Trong các thử nghiệm này, sẽ sử dụng dữ liệu từ tháng 1 năm 2009 trở đi. Các bộ phim có ít hơn 20 xếp hạng sẽ bị loại bỏ. Cách tạo bộ dữ liệu training và bộ dữ liệu thử nghiệm như sau:

Bộ dữ liệu training: xem xét dữ liệu từ tháng 1 năm 2009 đến tháng 3 năm

2014. Đối với mỗi khách hàng u_n, tạo một mẫu dữ liệu như đã đề cập trong phần 2.3. Điều này có nghĩa là đối với mọi xếp hạng phim, sẽ tạo một mẫu dữ liệu trong đó trình tự đầu vào chứa các xếp hạng trước đó cho đến thời điểm đó và nhãn thực là mã hóa duy nhất hoặc phần nhúng của phim được xếp hạng trong thời điểm thực tế. Hơn nữa chỉ tạo mẫu khi có ít nhất 5 xếp hạng phim trước xếp hạng thực tế và giới hạn số lượng xếp hạng trên mỗi mẫu là 100. Do đó, nếu khách hàng đã xếp hạng hơn 100 xếp hạng thì chỉ xem xét 100 xếp hạng mới nhất. Tập huấn luyện cuối cùng chứa 4053420 mẫu dữ liệu cho tổng số 32901 khách hàng khác nhau.

Bộ dữ liệu thử nghiệm: bộ dữ liệu này sẽ tập trung vào dự đoán xếp hạng của khách hàng từ tháng 4 năm 2014 cho đến tháng 4 năm 2015. Theo đó giai đoạn này là giai đoạn thử nghiệm. Mặc dù việc tối ưu hóa mô hình để tìm hiểu phim nào sẽ được xếp hạng tiếp theo (dự đoán ngắn hạn), tuy nhiên cũng đo lường cuối cùng khách hàng sẽ xếp hạng phim nào (dự đoán dài hạn). Do đó, đối với mỗi khách hàng, sẽ tạo ra các nhãn thực sự $y_n = (y_{n1}, y_{n2}, \dots, y_{nP})$ theo đó y_{nt} là bộ phim thứ t do khách hàng đánh giá u_n và P là số lượng phim được khách hàng đánh giá trong thời gian thử nghiệm. Để dự đoán các bộ phim được xếp hạng, tiếp tục tạo trình tự $x_n = (x_{n1}, x_{n2}, \dots, x_{nT})$, chứa những bộ phim mà khách hàng đã xem u_n trước thời gian kiểm tra. Sau đó giới hạn phim của chuỗi đầu vào trong 100 phim cuối cùng đánh giá của khách hàng. Dữ liệu kiểm tra cuối cùng chứa 3669 khách hàng khác nhau.

Trong bộ dữ liệu này bao gồm các master data sau:

- Movies: chưa danh sách các bộ phim thuộc nhiều thể loại khác nhau từ năm 2009 đến năm 2015, ví dụ: *"Toy Story (1995)", "Jumanji (1995)", "Tom and Huck (1995)",...*
- Genome_tags: bao gồm thông tin các nhãn được gắn cho các bộ phim, gồm: *"Adventure", "Action", "80s",...*
- Rates: chứa thông tin đánh giá của khách hàng cho từng bộ phim cụ thể, bao gồm nhiều bộ phim được đánh giá bởi nhiều khách hàng theo thang điểm từ 0 đến 5.

- Genome_scores: biểu diễn độ chính xác của việc gán nhãn cho từng bộ phim. Với thang điểm từ 0 đến 1.

Để tạo phép nhúng kết hợp với word2vec, bộ dữ liệu training sẽ được sử dụng. Đối với mỗi khách hàng, tạo một chuỗi các bộ phim mà khách hàng này đã xếp hạng từ tháng 1 năm 2009 đến tháng 3 năm 2014. Sau đó, các chuỗi này được đưa vào mô hình word2vec để tạo các mục nhúng.

Sau khi xử lý trước, thu được tổng cộng 10057 phim khác nhau. Trong trường hợp này, đánh giá tập dữ liệu này bằng tất cả các phương pháp được giải thích trong phần tiếp theo.

3.2. Cách thức thực nghiệm và đánh giá.

3.2.1. *Cách thức thực nghiệm.*

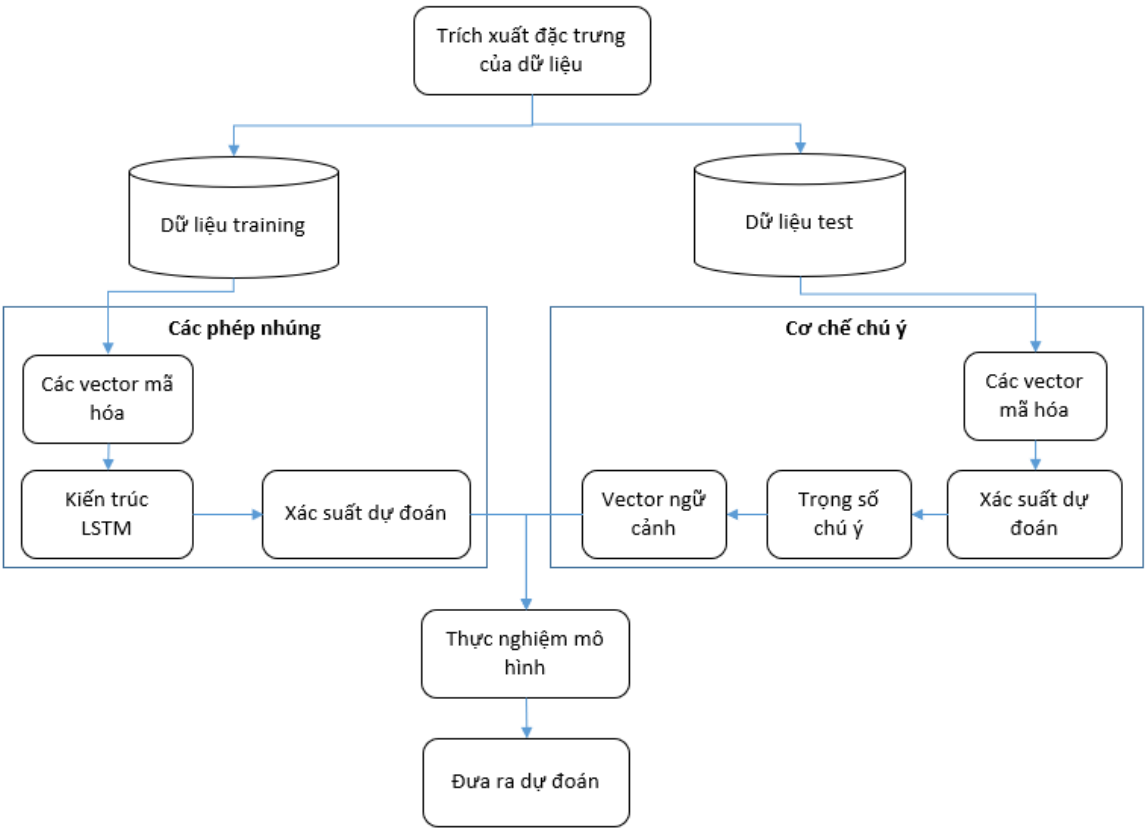
Trong luận văn này sẽ triển khai việc các phép nhúng và các cơ chế chú ý cho các mô hình bằng cách sử dụng bộ framework TensorFlow[1], cho phép tạo và đào tạo các mạng Nơron sâu mà không cần cung cấp các biểu thức của gradient, vì chúng được tính toán tự động. Đối với các mô hình mà các lần nhúng được tính bằng phương pháp Skip-gram, mã nguồn sử dụng gói python Gensim[33].

Để tìm ra các siêu tham số chính xác để huấn luyện các mô hình, và thử một tập hợp nhỏ các giá trị. Tuy nhiên, do thời gian cần thiết cho mỗi thử nghiệm với tập dữ liệu Movielens, báo cáo này không thực hiện tìm kiếm đầy đủ trong không gian tham số. Do đó, có khả năng là các siêu thông số được sử dụng không phải là tối ưu cho vấn đề này.

Để so sánh công bằng giữa các mô hình, số lượng tế bào thần kinh ẩn trong khối LSTM và kích thước nhúng (chỉ áp dụng trong bộ dữ liệu Movielens) đã được đặt thành các giá trị giống nhau giữa các mô hình khác nhau. Số lượng tế bào thần kinh ẩn được đặt thành 256 trong tập dữ liệu Movielens. Kích thước nhúng để thể hiện các phim khác nhau được đặt thành 64.

Đối với tất cả các mô hình, mạng được tối ưu hóa bằng phương pháp học tập Adam với kích thước lô là 128. Tốc độ học tập được đặt thành 0,001 cho cả hai tập dữ liệu. Phần còn lại của các tùy chọn cho phương thức Adam được đặt làm mặc

định trong gói Tensorflow. Tiếp theo đã thử bỏ phép học và chính quy hóa L2, nhưng nó không dẫn đến kết quả tốt hơn.



Hình 3.1. Phương pháp thực nghiệm đề xuất.

Để tránh mô hình sử dụng quá nhiều dữ liệu huấn luyện do quá nhiều thời gian huấn luyện, chúng ta đã sử dụng kỹ thuật Dừng sớm, trong đó chúng ta huấn luyện mô hình cho đến khi đánh giá trên một tập hợp xác nhận riêng biệt không được sử dụng để huấn luyện các điểm dừng để cải thiện. chúng ta cũng đã thử các khối RNN khác, chẳng hạn như GRU hoặc các biến thể LSTM khác nhau, nhưng chúng ta không có được độ chính xác tốt hơn. Cuối cùng, thử nghiệm việc xếp chồng nhiều hơn 1 lớp khối LSTM, nhưng không thành công để cải thiện hiệu suất. Tuy nhiên, đối với tập dữ liệu Movielens, việc thêm nhiều lớp hơn đã làm tăng thời gian đào tạo, do đó không thể đào tạo mô hình đủ lâu để xác định rằng không có cải tiến nào đạt được bằng cách thêm nhiều lớp hơn. Hơn nữa, chúng ta có thể cần phải điều chỉnh các siêu thông số khác cho hợp lý này.

3.2.2. *Cách thức đánh giá.*

Trong luận văn này sẽ tập trung đánh giá các dự đoán ngắn hạn (phim đầu tiên được xếp hạng trong giai đoạn thử nghiệm) và dự đoán dài hạn (tất cả các phim được xếp hạng trong giai đoạn thử nghiệm).

Một biện pháp dự đoán dài hạn được đưa ra trong luận văn này. Một số khách hàng chỉ đánh giá một vài bộ phim trong bộ dữ liệu test, trong khi những người khác có thể đánh giá hàng trăm bộ phim. Nếu cắt giảm ở mức 10, đối với một khách hàng đã xếp hạng 100 phim, thì chỉ có thể nhận được mức thu hồi tối đa là 0,1. Điều này làm cho sự đo lường không lý tưởng để biết mô hình tốt như thế nào. Do đó, nên sử dụng R-Precision làm thước đo. R-Precision giúp cắt giảm số lượng phim mà khách hàng đã đánh giá. Điều đó có nghĩa là giới hạn thay đổi tùy theo mỗi khách hàng. Nếu khách hàng xếp hạng 5 phim trong bộ dữ liệu test, tiếp theo xem xét 5 dự đoán hàng đầu, nếu khách hàng xếp hạng 100 phim trong bộ dữ liệu test, sử dụng việc xem xét 100 dự đoán hàng đầu. Do đó, nếu R là số lượng phim được xếp hạng trong khoảng thời gian thử nghiệm cho một khách hàng nhất định, sau đó, R-Precision được tính như sau:

$$R - Precision = \frac{\#\ positives\ top\ R}{R}$$

Điều đó làm cho phép đo sẽ nằm trong khoảng từ 0 đến 1 cho mọi khách hàng. Sau đó, tính trung bình số đo cho tất cả khách hàng.

3.3. Cài đặt phần mềm.

3.3.1. *Tổng quan phần mềm.*

Trong nghiên cứu này, sử dụng bộ framework Tensorflow, cho phép tạo và đào tạo các mạng Nơron thông qua việc tính toán tự động. Kèm theo đó sử dụng phương pháp Skip-gram thông qua việc sử dụng package Gensim. Phần tiếp theo sẽ trình bày chi tiết về hướng dẫn cài đặt phần mềm và chạy thử nghiệm.

3.3.2. *Cài đặt framework Tensorflow.*

Trong nghiên cứu này, sẽ sử dụng bộ framework Tensorflow có hỗ trợ GPU. Trước khi tiến hành cài đặt, môi trường cài đặt phần mềm có thông tin như sau:

- Window version: Window 10 Professional(64bit)
- RAM: 8GB
- Graphic driver: NVIDIA GeForce 840M
- Trình điều khiển GPU NVIDIA
- Bộ công cụ CUDA
- CUPTI
- cuDNN SDK 8.0.4
- Python phiên bản từ 3.5.x đến 3.8.x, lưu ý tensorflow chỉ hỗ trợ các phiên bản python này. Luận văn này sử dụng python 3.8.5.
- Phần mềm Anaconda quản lý các add on.

Sau khi đã đáp ứng đủ các yêu cầu phía trên, hệ thống đã sẵn sàng để cài đặt Tensorflow.

Đầu tiên, mở Anaconda Prompt (Anaconda3) chạy câu lệnh để tạo môi trường cho việc kiểm thử và đánh giá:

```
Conda create -n tensorflow-gpu python=3.8.5
```

Tiếp theo, sau khi đã tạo môi trường thành công, sử dụng câu lệnh sau để truy cập vào môi trường vừa tạo và tiến hành cài đặt Tensorflow:

```
Activate tensorflow-gpu
```

Tiếp theo chạy câu lệnh để cài đặt Tensorflow hỗ trợ GPU thông qua pip:

```
pip install tensorflow-gpu
```

Sau khi cài đặt xong, thực hiện câu lệnh sau để kiểm tra đã cài đặt thành công Tensorflow và đã nhận GPU hay chưa(Nếu như không có thông báo lỗi nào, thì việc cài đặt đã thành công):

```
python -c "import tensorflow as tf;
print(tf.reduce_sum(tf.random.normal([1000, 1000])))"
```

3.3.3. *Cài đặt package Gensim*

Tương tự như việc cài đặt Tensorflow, trước tiên cần truy cập vào môi trường vừa tạo khi cài đặt tensorflow, sau đó chạy câu lệnh sau để cài đặt Gensim:

```
pip install --upgrade genism
```

3.4. Thực nghiệm mã nguồn và đánh giá kết quả

3.4.1. *Thực nghiệm mã nguồn.*

Trong phần này, luận văn sẽ trình bày cách thiết lập mã nguồn để kiểm tra và đánh giá kết quả của các phép nhúng và áp dụng các cơ chế chú ý trong việc dự đoán hành vi khách hàng.

3.4.1.1. *Import các thư viện của Python.*

Như đã đề cập trong các phần trước, phần mã nguồn này sẽ dùng các thư viện đã được đề cập trong phần cài đặt đã nêu trước đó.

Cụ thể, mã nguồn sẽ sử dụng thư viện NumPy để xử lý dữ liệu tính toán ở mức độ cao, Pandas để phân tích dữ liệu và thực thi việc phân tích và Matplotlib để mô hình hóa dữ liệu.

```
import pandas as pd
import numpy as np
from gensim.models import Word2Vec
import random
from tqdm import tqdm
import matplotlib.pyplot as plt
%matplotlib inline
import warnings;
warnings.filterwarnings('ignore')
```

3.4.1.2. *Đọc dữ liệu nguồn.*

Như đã đề cập trước đó, bộ dữ liệu sẽ được sử dụng là Movielens, chứa hơn 20 triệu dòng dữ liệu đánh giá với hơn 465,000 dán nhãn của hơn 27 nghìn bộ phim đánh giá bởi hơn 138 nghìn người xem.

```
df_movies =pd.read_csv('movies.csv')
df_ratings =pd.read_csv('ratings.csv')
```

3.4.1.3. *Gộp dữ liệu và loại bỏ dữ liệu trống.*

Hợp nhất dữ liệu có nghĩa là kết hợp hai tập dữ liệu theo cách mà mỗi hàng trong cả hai tập dữ liệu đều căn chỉnh dựa trên các thuộc tính hoặc cột chung. Ở đây,

chúng ta sẽ hợp nhất bộ dữ liệu phim và xếp hạng để lấy ID phim, ID khách hàng và tiêu đề phim trong một khung dữ liệu. Và loại bỏ những dòng dữ liệu có chứa các giá trị trống.

```
df =pd.merge(df_movies,df_ratings)
df.dropna(inplace=True)
```

3.4.1.4. *Tiền xử lý dữ liệu.*

Mục đích của việc này là làm mịn dữ liệu trước khi tiến hành phân tích. Đầu tiên, ID sẽ được thay đổi định dạng sang kiểu chuỗi và loại bỏ những UserID trùng nhau.

```
df['movieId']=df['movieId'].astype(str)
users =df["userId"].unique().tolist()
len(users)
```

Sau khi làm mịn ta thu được hơn 162 nghìn dữ liệu của khách hàng, mỗi khách hàng đề có lịch xử xem phim riêng.

3.4.1.5. *Phân tách dữ liệu.*

Để đáp ứng việc thử nghiệm performance của các mô hình, dữ liệu ban đầu cần được tách và sử dụng theo công thức 90% dữ liệu training và 10% dữ liệu thử nghiệm.

```
random.shuffle(users)
# extract 90% of user ID's
users_train =[users[i] fori in
range(round(0.9*len(users)))]
# split data into train and validation set
train_df =df[df['userId'].isin(users_train)]
validation_df =df[~df['userId'].isin(users_train)]
```

3.4.1.6. *Chiến lược thực hiện.*

Trong phần này, mã nguồn sẽ được thay đổi tuy theo phép nhúng hay cơ chế chú ý nào sẽ được áp dụng, trong mã nguồn phía dưới, mục đích để đưa ra các gợi ý về các bộ phim nên xem cho người xem sử dụng phép nhúng RNN-Baseline:

```
#list to capture watch history of the users
```

```
watch_train = []
# populate the list with the movie ID
for i in tqdm(users_train):
    temp = train_df[train_df["userId"] ==
i]["movieId"].tolist()
    watch_train.append(temp)
```

3.4.1.7. *Đào tạo cho các mô hình.*

Để đào tạo các mô hình, mã nguồn sẽ sử dụng thư viện gensim. Module này chứa bộ thư viện Word2Vec, được sử dụng để xử lý lượng dữ liệu lớn và hỗ trợ giao diện Pythonic, mã nguồn bên dưới áp dụng RNN-emb-Word2vce:

```
model = Word2Vec(window = 10, sg = 1, hs = 0,negative =
10,
alpha=0.03, min_alpha=0.0007, seed = 14)
model.build_vocab(watch_train, progress_per=200)
model.train(watch_train, total_examples =
model.corpus_count, epochs=10, report_delay=1)
```

Tiếp theo có thể in ra mô hình để kiểm tra nếu cần.

```
X = model[model.wv.vocab]
X.shape
```

3.4.2. *Đánh giá kết quả.*

Trong phần đánh giá này, luận văn sẽ làm rõ hai kết quả đã cho mục tiêu được đặt ra trong chương 1. Thứ nhất, chúng ta có thể thấy rằng thông qua việc sử dụng các phép nhúng kết hợp với word2vec sẽ giúp cho việc tối ưu hóa vector dữ liệu tương tác của khách hàng đề phục vụ cho việc dự đoán hành vi trong tương lai. Trong các ví dụ trong phần 3.4.2.1, chúng ta sẽ thấy rõ hơn việc các phép kết hợp với word2vec sẽ kết hợp thông tin về thời gian phát hành bộ phim, thể loại hoặc một chuỗi các sự kiện cụ thể. Điều này có nghĩa là có thể có một số khách hàng xem liên tiếp các bộ phim cùng thời kỳ, cùng thể loại hoặc cùng một câu chuyện.

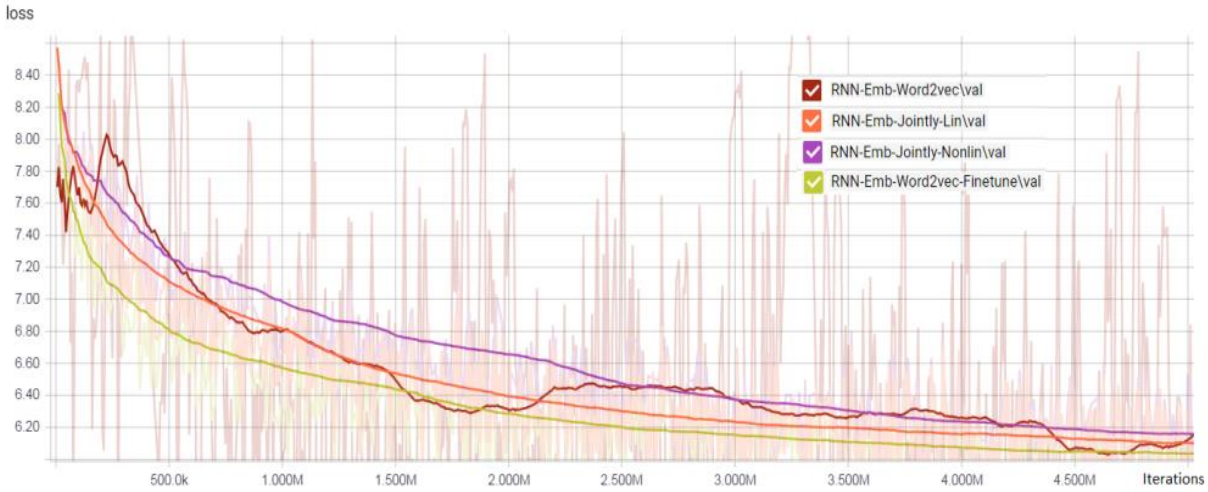
Thứ hai, các cơ chế chú ý sử dụng các trọng số chú ý đã được thử nghiệm góp

phần đưa ra các dự đoán hành vi của khách hàng trong tương lai dựa vào các tương tác trong quá khứ. Trong phần phân tích tại mục 3.4.2.2 sẽ cho chúng ta thấy việc dự đoán hành vi của khách hàng trong tương lai cho bộ dữ liệu cụ thể đã đề cập ở phần trước và hỗ trợ việc đưa ra các chiến lược phù hợp để tiếp cận khách hàng tốt hơn.

Cần lưu ý rằng việc dự đoán hành vi khách hàng và việc giới thiệu phim tới khách hàng có một điểm chung là đưa ra các dự đoán của khách hàng trong tương lai để từ đó có những cơ chế quảng cáo, tiếp thị và kinh doanh phù hợp, tránh đi ngược lại xu hướng của khách hàng. Tuy nhiên có một điểm khác biệt giữa hai hình thức này là phân tích hành vi khách hàng mang ý nghĩa dự đoán rộng và phức tạp hơn so với giới thiệu phim tới khách hàng.

3.4.2.1. *Đánh giá kết quả việc phân tích phép nhúng*

Trong phần này, luận văn sẽ trình bày kết quả phân tích cho các phép nhúng thu được bằng các phương pháp khác nhau. Mục đích là để quan sát các đặc điểm nào được thể hiện trong các mục nhúng. Lưu ý rằng khi tham khảo RNN-Emb-Word2vec, Skip-gram với lấy mẫu gián tiếp (word2vec) được sử dụng trực tiếp, với phép nhúng RNN-Emb-Jointly-Lin và RNN-Emb-Word2vec-Finetune phép nhúng ma trận W_{emb}, và với phép nhúng RNN-Emb-Jointly-Nonlin sẽ sử dụng phép toán đã được giới thiệu tại chương 2 là $\tanh(W_{emb}x_t + b_{emb})$.



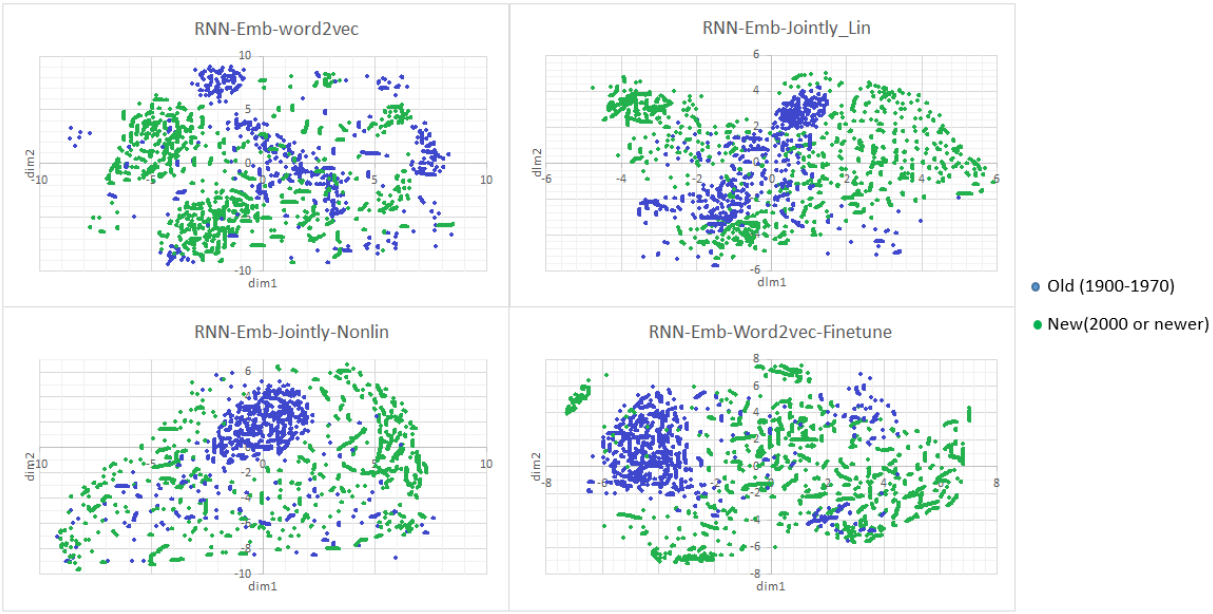
Hình 3.2. Độ mất mát của bộ xác thực của các phương pháp nhúng khác nhau. Trục x đại diện cho số lần lặp lại.

Bảng 3.1. Năm bộ phim tương tự hàng đầu.

	RNN-Emb-Word2vec	RNN-Emb-Jointly-Lin	RNN-Emb-Jointly-Nonlin	RNN-Emb-Word2vec-Finetune
Finding Nemo	Monsters, Inc. Incredibles, The Shrek Toy Story 2, Pirates of the Caribbean	Incredibles, The Green Mile, The Monsters, Inc. Ocean's Eleven Pirates of the Caribbean	Very Potter Musical Pentagon Wars Neil Young Devdas Find Me Guilty	Incredibles, The Monsters, Inc. Ocean's Eleven Pirates of the Caribbean Shrek
Star Wars IV	Star Wars: Episode V Star Wars: Episode VI Raiders of the Lost Ark Matrix, The Silence of the Lambs	Fugitive, The Teenage Mutant Ninja Tur... Richard Pryor: Live in... Alone in the Wilderness Raiders of the Lost Ark	Terminator 2 Wallace & Gromit Spirited Away Ponterosa Star Wars VI	Star Wars: Episode VI Silence of the Lambs Star Wars: Episode I Terminator, The Shawshank Redemption
[REC]	Orphanage, The Devil's Backbone Martyrs Eden Lake Descent, The	Inside High Tension Martyrs Tale of Two Sisters, A Frailty	Splinter Turtles Can Fly Thesis Dark Water Henry & June	Orphanage, The Devil's Backbone 3 Extremes Battle Royale Grudge 3, The
Shining, The	Clockwork Orange, A Full Metal Jacket Taxi Driver Trainspotting Psycho	Jaws Clockwork Orange, A Trainspotting Truman Show, The World According to Mons...	World According to Mons... The Devils Wal-Mart: The High Cost... Trainspotting Stand by Me	Stand by Me Psycho Clockwork Orange, A Trainspotting Repulsion

Trong ví dụ đầu tiên, chúng ta thấy khi người xem bộ phim "*Finding Nemo*" thì ngay lập tức sẽ có 5 bộ phim giống nhau nhất với bộ phim này được liệt kê và gợi ý thêm cho khách hàng. Chúng ta thấy rằng tất cả các phương pháp ngoại trừ RNN-Emb-Jointly-Nonlin đều trả về các bộ phim có cùng thể loại, bao gồm các bộ phim dành cho thanh thiếu niên khác trong top 5, như *Monsters, Inc or The Incredibles*. Ngược lại, với phép nhúng RNN-Emb-Jointly-Nonlin khó tìm thấy những điểm tương đồng với "*Finding Nemo"* được gợi ý cho người xem trong top 5 phim giống

nhau nhất. Ví dụ thứ hai cho thấy 5 bộ phim được gợi ý cho người xem cùng chủ đề với phim "*Star War IV*". Các mô hình RNN-Emb-Word2vec và RNN-Emb-Word2vec-Finetune dường như mã hóa mật thiết các bộ phim có liên quan rất nhiều về mặt ngữ nghĩa, vì chúng có nhiều hơn một phim *Star War* trong top 5. So với trường hợp trước, chúng ta thấy các mối quan hệ trong một số phim thuộc top 5 với phép nhúng RNN-Emb-Jointly-Nonlin, giống như phim *Star War* hoặc các phim Sci-Fi như *Terminator 2*. Trong ví dụ thứ ba, ba mô hình dường như đặt các bộ phim *Hornor* / *Drama* khác gần với [REC]. Điều thú vị, trong ví dụ cuối cùng, 2 bộ phim giống nhau nhất đang dẫn đầu The Shining cho RNN-Emb-Word2vec là hai bộ phim của cùng một đạo diễn.

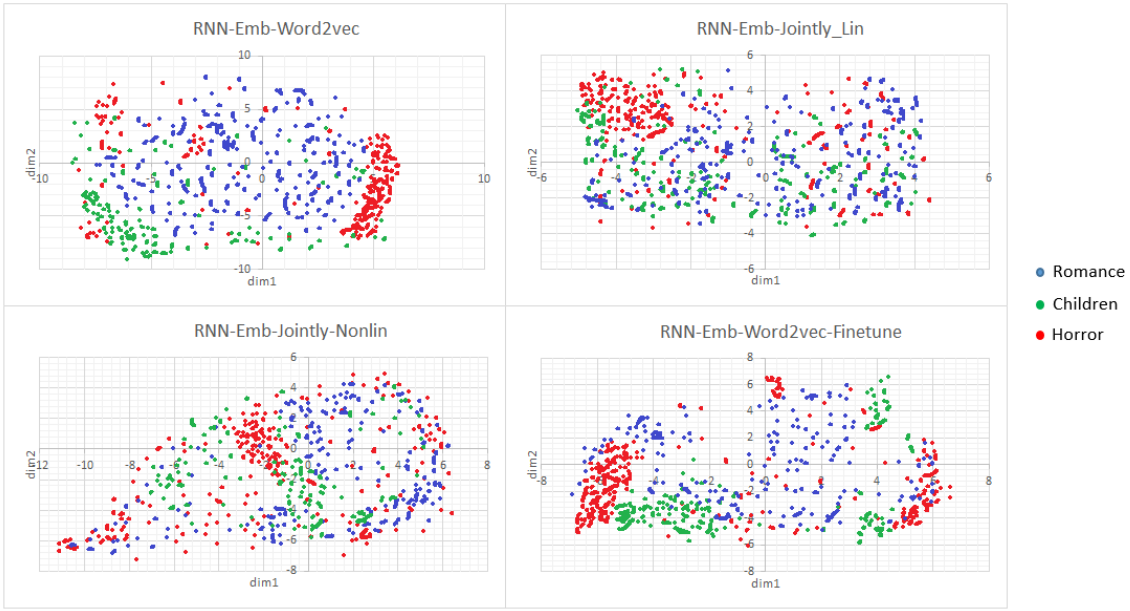


Hình 3.3. Biểu diễn phép nhúng cho việc phân tích phim cũ và phim mới.

Hình 3.9 cho thấy kết quả của các phép nhúng cho các bộ phim cũ (từ năm 1900 đến năm 1970) và phim mới (từ năm 2000 trở đi) sau khi giảm kích thước của các phép nhúng xuống giá trị là 2 với t-SNE. Đối với tất cả các mô hình, cho thấy một số khu vực có mật độ phim cũ tập trung cao.

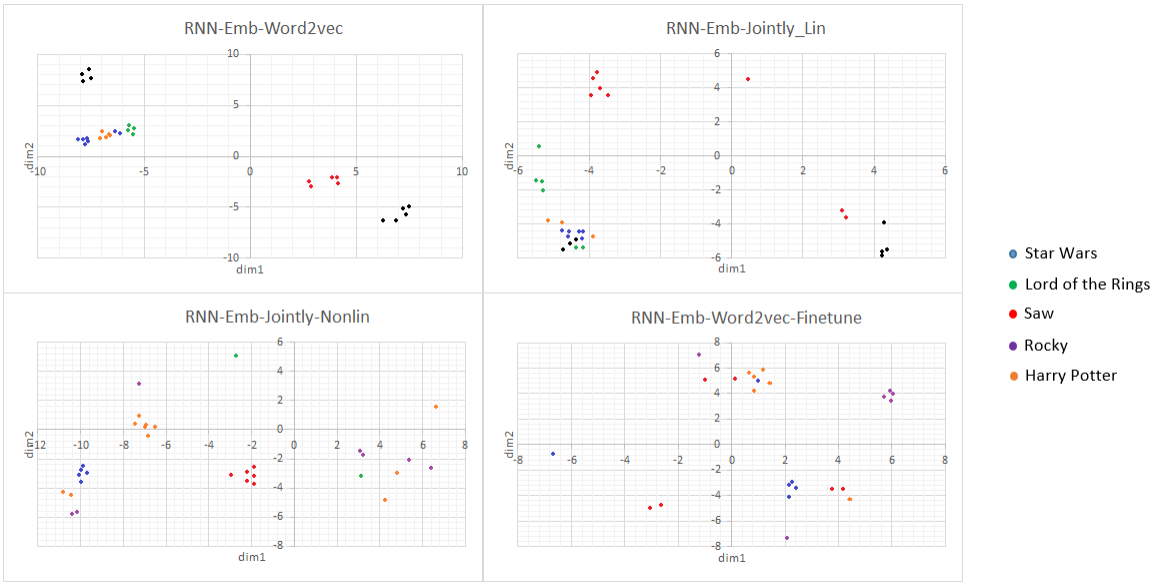
Hình 3.4. Biểu diễn phép nhúng cho các bộ phim được phát hành vào các năm khác nhau.

Trong hình 3.10, chúng ta tiếp tục xem xét việc các phép đại diện nhúng của các phạm vi cụ thể hơn của năm phát hành. Trong cả ba phương pháp, chúng ta thấy các cụm tập trung theo năm phát hành. Điều thú vị là trong các mô hình có sử dụng phép nhúng kết hợp, chúng ta có thể nhận thấy rằng các nhóm phim kết thúc cùng thời gian (2010-2013, 2013-2014, 2014 hoặc mới hơn) được đặt gần nhau, trong khi đối với các mô hình khác, các nhóm này xa hơn.



Hình 3.5. Biểu diễn phép nhúng cho các bộ phim khác nhau, khác thể loại.

Hình 3.11 cho thấy việc biểu diễn các phép nhúng cho các bộ phim thuộc các thể loại khác nhau. Chúng ta có thể thấy rằng tất cả các phương pháp đều có thể gộp các bộ phim kinh dị lại với nhau. Chúng ta thấy một nhóm lớn phim kinh dị và các nhóm thể loại nhỏ khác, có thể là các loại phụ bản cụ thể của phim kinh dị. Các mô hình sử dụng phép nhúng dường như cũng nhóm một số phim dành cho trẻ em lại với nhau, đặc biệt phép nhúng RNN-Emb-Word2vec và phép nhúng RNN-Emb-Word2vec_Finetune. RNN-Emb-Word2vec có sự phân tách rõ ràng giữa các thể loại dành cho trẻ em và thể loại kinh dị, vì hai thể loại khá khác nhau. Ngược lại, các mô hình sử dụng phép nhúng kết hợp lại biểu diễn một số khu vực có sự trộn lẫn giữa phim trẻ em và phim kinh dị. Có rất nhiều vùng khác mà chúng ta có thể thấy thuộc thể loại phim tình cảm, nhưng có thể là vì thể loại phim tình cảm là một thể loại rộng và liên quan đến nhiều thể loại khác.



Hình 3.6. Biểu diễn phép nhúng cho các bộ phim khác nhau.

Hình 3.12 cho thấy phép nhúng của một số đoạn phim đặc biệt. Như chúng ta thấy, phép nhúng RNN-Emb-Word2vec mã hóa chặt chẽ tất cả các thể loại khác nhau, ngoại trừ bộ phim *Rocky* hoàn toàn tách biệt so với các bộ phim khác. Có vẻ như ba phương pháp còn lại gộp nhóm một số phim của mỗi thể loại, nhưng chúng ta có thể thấy rằng mỗi số phim về thể loại khác nhau nằm trong một khu vực khác nhau.

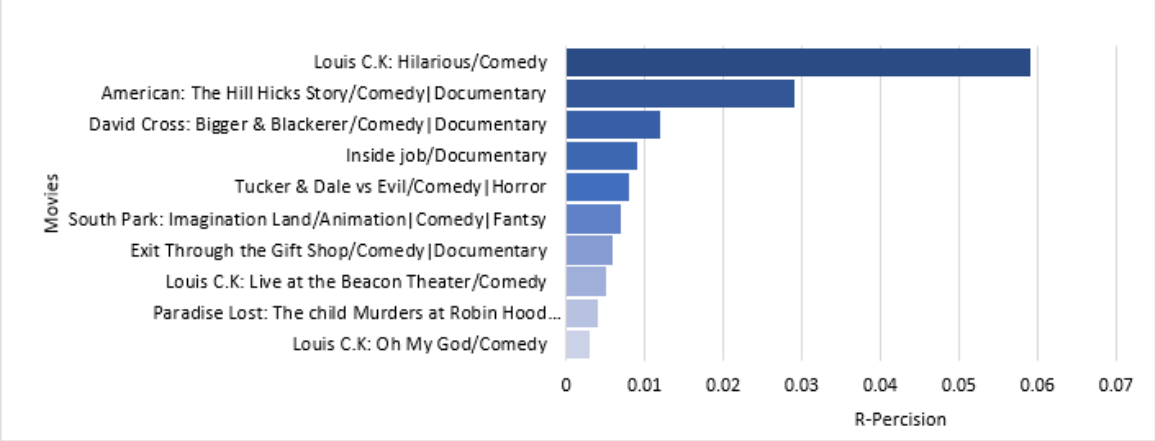
Nói chung, chúng ta có thể quan sát thấy rằng tất cả các biểu diễn phép nhúng chứa một số thông tin cơ bản về các mục đang mã hóa. Word2vec học các biểu diễn thông qua việc sử dụng ngữ cảnh của các mục. Như chúng ta đã thấy, phép nhúng sử dụng word2vec dường như kết hợp thông tin về thời gian phát hành bộ phim, thể loại hoặc thời gian cụ thể. Điều này có nghĩa là có thể có một số khách hàng xem liên tiếp các bộ phim cùng thời kỳ, cùng thể loại hoặc cùng một câu chuyện tiếp nối nhau. Chúng ta đã thấy rằng các phương pháp khác cũng có thể kết hợp loại thông tin này trong các phép nhúng. Các phương pháp này cố gắng tối ưu hóa biểu diễn phép nhúng cho từng nhiệm vụ cụ thể, trong đó RNN-Emb-Word2vec-Finetune bắt đầu với việc biểu diễn phép nhúng sử dụng word2vec. Vì nhiệm vụ bao gồm dự đoán các bộ phim được đưa ra dựa vào các bộ phim trước đây, các cách biểu diễn phép nhúng cũng được tập trung đào tạo và có kết hợp với việc tính toán bối cảnh.

3.4.2.2. *Phân tích tính tập trung*

Như đã đề cập, trọng số chú ý có thể cung cấp mức độ tập trung của mô hình vào từng phần tử của chuỗi để đưa ra dự đoán. Do đó, trong số chú ý sẽ rất hữu ích để giải thích các dự đoán. Trong phần này, một số ví dụ về các dự đoán và trọng số chú ý cho các phương pháp khác nhau sẽ được đưa ra và đánh giá.

Đầu tiên, phép nhúng có trọng số tuyến tính RNN-Att-Emb-Lin sẽ được sử dụng để làm ví dụ, vì nó là phương pháp cung cấp khả năng diễn giải cao hơn cho một phép nhúng.

Fahrenheit 9/11	Crumb	Louis C.K: Shameless	Louis C.K: Chewed Up
Forgetting Sarah Marshall	White Diamond	Waiting for Guffman	Breakfast
Style War	Ginger Snaps	Bride of Re-Animation	All Watched Over by Machines
Perfume: The Story of Murder	Mr. Nobody	Inocence	Party Monster
Crazies	Yes Men Fix the World	Four Kind	Red State
Crazy, Stupid, Love	Wedding Planner	Another Earth	Crazies
Cabin in the Wood	Black Swan	Jenifer's Body	Ameraca Pie Presidents



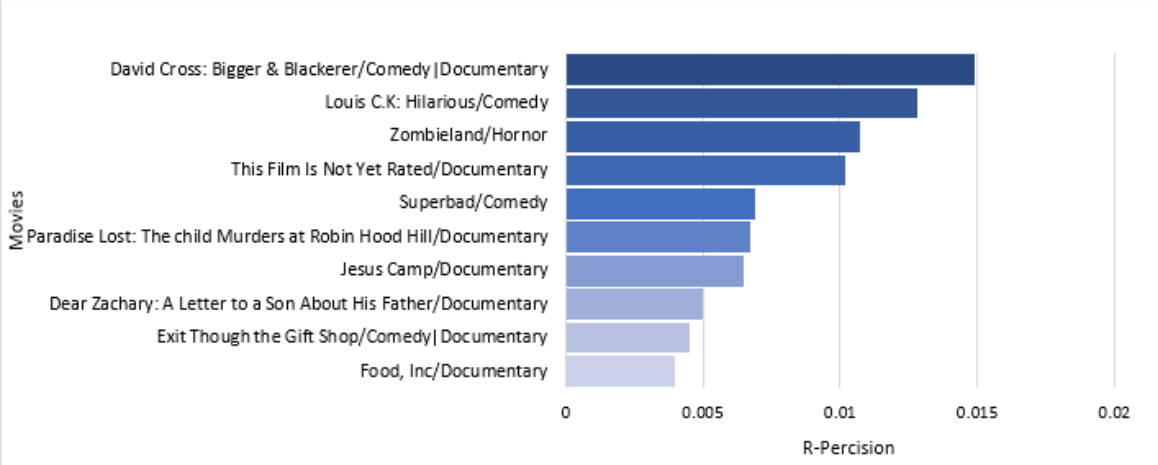
Hình 3.7. Cơ chế chú ý kết hợp phép nhúng (tuyến tính) - RNN-Att-Emb-Lin.

Hình 3.13 cho thấy ví dụ đầu tiên. Ở trên cùng, chúng ta thấy danh sách 28 bộ phim theo trình tự từ thấp lên cao được khách hàng quan tâm nhất. Thứ tự đi từ trái sang phải, từng hàng một, hàng cuối cùng đại diện cho những bộ phim mới nhất. Trọng số chú ý được biểu thị bằng màu sắc, trong đó màu tối hơn cho thấy mô hình tập trung nhiều chú ý hơn. Ở phía dưới, chúng ta thấy 10 bộ phim được dự đoán sẽ đứng đầu xu hướng và sẽ được người xem ưu tiên lựa chọn. Cụ thể, ta thấy rằng phép nhúng đã dự đoán mạnh mẽ cho show truyền hình "*Louis CK*" có khả năng được xem cao nhất với xác suất được xếp hạng cao nhất. Tiếp đó, ta có thêm một bộ

nữa hiển thị trong 10 xác suất dự đoán hàng đầu và xác suất dự đoán thứ hai là "*David Cross: Bigger & Blackerer"* đó là một chương trình hài kịch khác. Với phép nhúng tuyến tính RNN-Att-Emb-Lin ta thu được dự đoán với xác suất cao tập trung vào thể loại hài kịch.

Hình 3.14 cho thấy các kết quả sự chú ý đến đối với phép nhúng có trọng số phi tuyến tính (RNN-Att-Emb-Nonlin). Qua các kết quả này, chúng ta nhanh chóng nhận thấy rằng trọng số chú ý được phân bổ nhiều hơn giữa các bộ phim khác nhau.

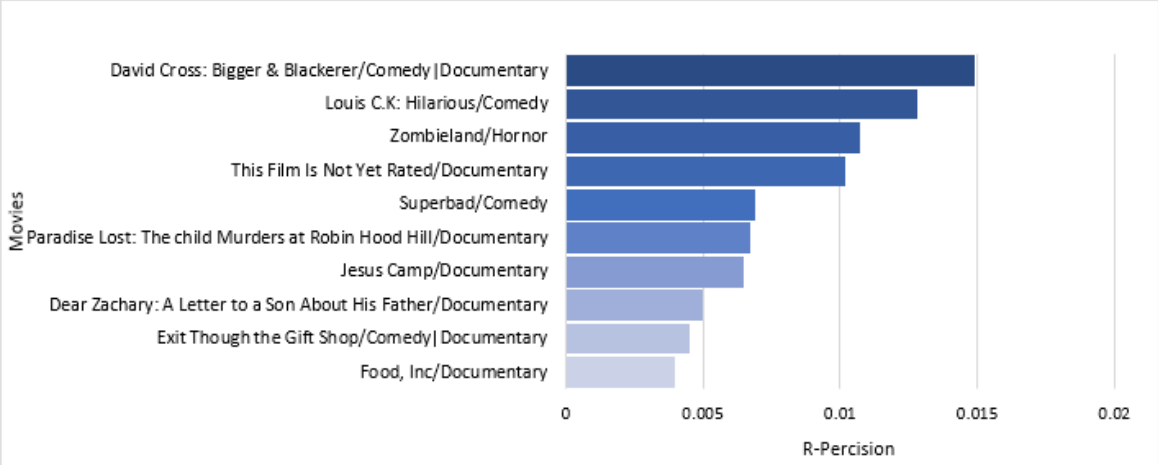
Fahrenheit 9/11	Crumb	Louis C.K: Shameless	Louis C.K: Chewed Up
Forgetting Sarah Marshall	White Diamond	Waiting for Guffman	Breakfast
Style War	Ginger Snaps	Bride of Re-Animation	All Watched Over by Machines
Perfume: The Story of Murder	Mr. Nobody	Inocence	Party Monster
Crazies	Yes Men Fix the World	Four Kind	Red State
Crazy, Stupid, Love	Wedding Planner	Another Earth	Crazies
Cabin in the Wood	Black Swan	Jenifer's Body	Ameraca Pie Presidents



Hình 3.8. Cơ chế chú ý kết hợp nhúng (phi tuyến tính) - RNN-Att-Emb-Nonlin.

Trong ví dụ 3.14, mô hình dự đoán chỉ một chương trình giải trí *Louis CK* nằm trong top 10 trong ví dụ đầu tiên, trái ngược với ba dự đoán bởi sự chú ý đến các phép nhúng có trọng số tuyến tính, có thể là do hiện tại mô hình tập trung vào các bộ phim khác nhau.

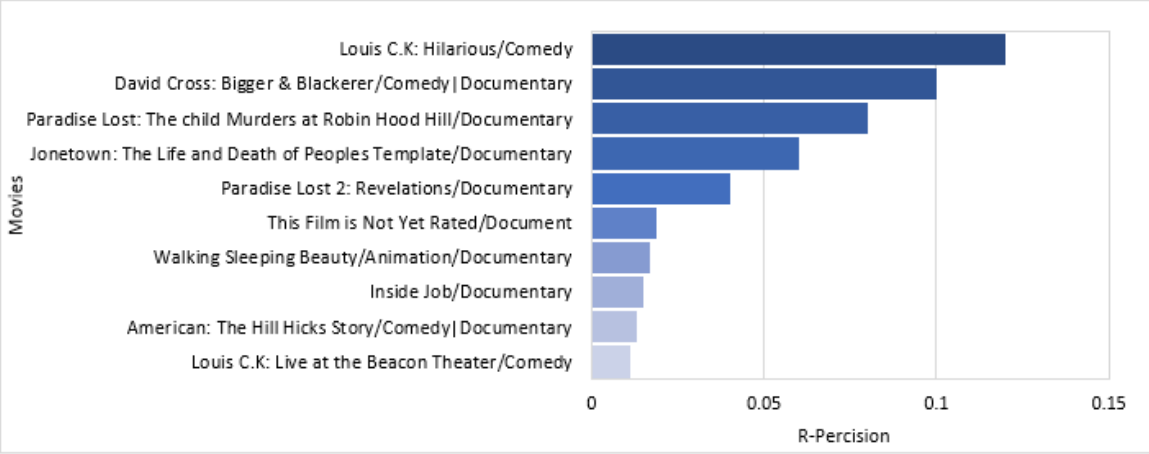
Fahrenheit 9/11	Crumb	Louis C.K: Shameless	Louis C.K: Chewed Up
Forgetting Sarah Marshall	White Diamond	Waiting for Guffman	Breakfast
Style War	Ginger Snaps	Bride of Re-Animation	All Watched Over by Machines
Perfume: The Story of Murder	Mr. Nobody	Inocence	Party Monster
Crazies	Yes Men Fix the World	Four Kind	Red State
Crazy, Stupid, Love	Wedding Planner	Another Earth	Crazies
Cabin in the Wood	Black Swan	Jenifer's Body	Ameraca Pie Presidents



Hình 3.9. Cơ chế chú ý kết hợp phép nhúng trên trạng thái ẩn (phi tuyến tính) - RNN-Att-HS-Nonlin.

Hình 3.15 cho thấy các ví dụ về cơ chế chú ý trạng thái ẩn với trọng số phi tuyến tính (RNN-Att-HS-Nonlin). Ở đây chúng ta cần lấy xem xét rằng trạng thái ẩn mã hóa thông tin từ các đầu vào trước đó. Do đó, thực tế là mô hình đang tập trung vào một trạng thái ẩn cụ thể không thể đảm bảo rằng mô hình đang tập trung vào một bộ phim cụ thể. Tuy nhiên, nếu đầu vào của một phim khiến dân mạng phải chú ý mạnh mẽ đến trạng thái ẩn, nó có thể là một tín hiệu cho thấy bộ phim cụ thể này có ảnh hưởng mạnh mẽ đến phần cuối cùng sự dự đoán. Trong trường hợp này, chúng ta thấy các điểm tương đồng với sự chú ý đến các nhúng với trọng số phi tuyến tính, vì sự chú ý được chia đều trong một số bộ phim. Điều đó có vẻ để dẫn đến sự đa dạng hơn trong các dự đoán. Ví dụ, chỉ có một chương trình “Louis C.K.” hiển thị trong 10 dự đoán hàng đầu trong hình 3.15 đầu tiên.

Fahrenheit 9/11	Crumb	Louis C.K: Shameless	Louis C.K: Chewed Up
Forgetting Sarah Marshall	White Diamond	Waiting for Guffman	Breakfast
Style War	Ginger Snaps	Bride of Re-Animation	All Watched Over by Machines
Perfume: The Story of Murder	Mr. Nobody	Inocence	Party Monster
Crazies	Yes Men Fix the World	Four Kind	Red State
Crazy, Stupid, Love	Wedding Planner	Another Earth	Crazies
Cabin in the Wood	Black Swan	Jenifer's Body	Ameraca Pie Presidents



Hình 3.10. Cơ chế chú ý kết hợp phép nhúng trên trạng thái ẩn (tuyến tính) dựa trên RNN-Att-HS-Lin.

Hình 3.16 hiển thị các ví dụ cho cơ chế chú ý với sự chú ý đến trạng thái ẩn với trọng số tuyến tính (RNN-Att-HS-Lin). Chúng ta quan sát thấy rằng mô hình dường như có xu hướng tập trung mạnh mẽ ở trạng thái ẩn cuối cùng, điều này xảy ra trong nhiều ví dụ. Trong trường hợp này, thật khó biết nếu mô hình đang biết rằng các phim cuối cùng quan trọng hơn và sử dụng các phim này cho các dự đoán, hoặc nếu mô hình đang mang thông tin của toàn bộ chuỗi đến các trạng thái ẩn cuối cùng. Do đó, trong trường hợp này, không thể xác định rằng trọng số chú ý là hữu ích để phát hiện phim nào quan trọng hơn trong dự đoán. Nói chung, chúng ta nhận thấy rằng việc sử dụng sự chú ý đến các nhúng với trọng số tuyến tính cung cấp một cái nhìn sâu sắc về lý do tại sao mô hình dự đoán một bộ phim là bộ phim có xác suất cao nhất. Phương pháp này cung cấp một số bộ phim là quan trọng nhất đối với mô hình mà có những điểm tương đồng với các bộ phim được dự đoán hàng đầu. Các mô hình có trọng số chú ý phi tuyến tính cung cấp các bộ phim lớn hơn, quan trọng để mô hình đưa ra dự đoán. Chúng ta nhận thấy rằng điều này giúp đa dạng hơn trong các dự đoán, mặc dù việc có nhiều phim có cùng mức độ chú ý có thể làm cho mô

hình ít hữu ích hơn cho mục đích giải thích. Chúng ta tin rằng có nhiều dự đoán hơn có thể là lý do tại sao các phương pháp phi tuyến tính đạt được độ R-Precision tốt hơn. Chỉ xem xét một hoặc một vài bộ phim ngụ ý rằng các dự đoán hàng đầu đều liên quan nhiều đến những bộ phim này. Ở một số khách hàng, đây có thể là những dự đoán tốt trong ngắn hạn, vì họ có thể quan tâm đến những bộ phim liên quan đến những bộ phim cuối cùng họ đã xem. Ví dụ: nếu khách hàng xem một bộ phim liên quan đến Artificial Intelligence, người xem có thể quan tâm đến các bộ phim khác về chủ đề này trong thời gian ngắn hạn. Tuy nhiên, sau một số tháng quan tâm đến chủ đề này có thể giảm đi. Top 10 dựa trên loại phim có nhiều khả năng có nhiều kết quả trùng khớp với phim đã xem thực cho khách hàng hơn top 10 chỉ dựa trên các phim liên quan đến Trí tuệ nhân tạo. Cuối cùng, mô hình với sự chú ý tuyến tính đến trạng thái ẩn thường tập trung hầu hết sự chú ý của nó vào trạng thái ẩn cuối cùng. Có hai trường hợp có thể giải thích hành vi này. Đầu tiên là rằng mô hình đang thực sự tập trung vào những bộ phim này, vì những bộ phim cuối cùng có thể là quan trọng để xác định những bộ phim mà khách hàng sẽ quan tâm. Thứ hai là mô hình chỉ mang tất cả thông tin từ các bộ phim trước, vì vậy phần cuối cùng bị ẩn trạng thái chứa thông tin về tất cả các bộ phim trước đó. Trong trường hợp cuối cùng này, không thể sử dụng trọng số chú ý để biết đâu là phim quan trọng nhất những dự đoán.

KẾT LUẬN VÀ ĐỊNH HƯỚNG NGHIÊN CỨU

Luận án này nghiên cứu việc sử dụng mạng thần kinh hồi quy để dự đoán hành vi của khách hàng từ dữ liệu tương tác trong quá khứ. RNN đã chứng tỏ tình hữu dụng trong việc lập mô hình dữ liệu tương tác của khách hàng để dự đoán các mặt hàng mà khách hàng sẽ mua trong tương lai. Việc sử dụng kỹ thuật nhúng đã cải thiện đáng kể hiệu năng mô hình RNN đơn giản trong các dự đoán ngắn hạn và dài hạn.

Các phương pháp nhúng khác nhau có thể biểu diễn các đặc trưng khác nhau mà khách hàng quan tâm như thời gian bộ phim được phát hành, thể loại hoặc câu chuyện. Kỹ thuật Word2vec có ưu thế rõ ràng hơn một chút, nhưng các phương pháp khác có thể đạt được hiệu suất tương tự với mục tiêu này. Nghiên cứu trong tương lai sẽ là khám phá các phương pháp để học cách nhúng bằng cách sử dụng dữ liệu mục (ví dụ như năm của bộ phim, đạo diễn, v.v.) hoặc sử dụng phép nhúng biểu đồ tri thức.

Cơ chế chú ý là một công cụ hữu ích để giải thích các dự đoán được thực hiện bởi các mô hình. Trong các thử nghiệm với tập dữ liệu Movielens, sự chú ý đến các nhúng có trọng số tuyến tính cung cấp một bộ phim hoặc một vài bộ phim quan trọng nhất đối với mô hình để đưa ra dự đoán. Đặc biệt kỹ thuật này cho phép cải thiện hiệu suất dự đoán dài hạn.

Một lĩnh vực nghiên cứu quan trọng sẽ là làm thế nào để biểu diễn thời gian điện tử trong RNN. RNN có thể có ưu thế để biểu diễn dữ liệu tuần tự như trong các nghiên cứu về xử lý ngôn ngữ tự nhiên, nơi mà khái niệm thời gian không tồn tại (không có thời gian giữa các từ, chúng chỉ thường xuyên được lấy mẫu lần lượt). Tuy nhiên, trong các nhiệm vụ như biểu diễn các tương tác của khách hàng, có một khoảng thời gian giữa mỗi lần tương tác. Thời gian này có thể chứa thông tin quan trọng, vì không giống nhau nếu một lần tương tác xảy ra sau lần trước đó vài giây so với lần tương tác xảy ra sau một năm. Chúng ta đã thấy rằng RNN không chứa thông tin này và chúng chỉ có thể đại diện cho thứ tự của chuỗi các tương tác.

TÀI LIỆU THAM KHẢO

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. “A Nơron attention model for abstractive sentence summarization”. *EMNLP* (2015).
[2] A. Graves, A. Mohamed, and G. Hinton. “Speech recognition with deep recurrent Nơron networks”. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference* (2013), pp. 6645–6649.
[3] A. Graves. “Generating sequences with recurrent Nơron networks”. *Arxiv preprintarXiv:1308.0850*(2013).
[4] A. Graves and J. Schmidhuber. “Framewise phoneme classification with bidirectional lstm and other Nơron network architectures”. *IEEE Transactions on Neural Networks* 18(5) (2005), pp. 602–610.
[5] Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. “Session-based Recommendations With Recurrent Neural Networks”. *ICLR* (2016).
[6] B. Carterette. “Multiple testing in statistical analysis of systems-based information retrieval experiments”. *ACM Transactions on Information Systems (TOIS)*
[7] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult”. *IEEE Transactions on Neural Networks* 5 (1994), pp. 157–166.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn/luận án qua phần mềm DoIT một cách trung thực và đạt kết quả mức độ tương đồng % toàn bộ nội dung luận văn/luận án. Bản luận văn/luận án kiểm tra qua phần mềm là bản cứng đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học Viện.

Hà Nội, ngày tháng năm
HỌC VIÊN CAO HỌC/NCS
(Ký và ghi rõ họ tên)