Phân tích cảm xúc khách hàng qua các bình luận trên trang web bán hàng

MỞ ĐẦU

1. Bối cảnh (thêm lí do)
2. Một số giải pháp hiện tại và giải pháp đề xuất
3. Những đóng góp và kết quả đạt được
4. Nội dung của khóa luận

Ngày nay, với sự gia tăng về số lượng và nội dung các trang thương mại điện tử như lazada.vn, tiki.vn, adayroi.vn và các mạng xã hội như Facebook, Twitter,… Internet không chỉ còn là nơi cung cấp thông tin ở dạng sự kiện mà còn là nơi người dùng bày tỏ cảm xúc, trao đổi cảm nhận, kinh nghiệm về các vấn đề đời sống hoặc về các vấn đề mà chính người dùng đó đang gặp phải. Có rất nhiều các nghiên cứu trong việc khai phá nội dung văn bản để tìm ra những điều mới mẻ phục vụ kinh doanh. Chẳng hạn như nhu cầu tổng hợp những phản hồi hay bình luận của người dùng trên Internet để đưa ra kết luận một sản phẩm có tốt hay không? Phản ứng của khách hàng đối với một sự kiện nào đó như thế nào?…  
Đối với lĩnh vực khai phá dữ liệu truyền thống, thay vì tập trung vào lịch sử người dùng như lịch sử mua bán, thời gian truy cập…, lĩnh vực khai phá cảm xúc người dùng lại tập trung vào việc phân tích ý nghĩa của các bình luận trên các trang bán hàng, thông tin hay mạng xã hội. Do đó, có thể hiểu bài toán phân loại cảm xúc người dùng là sự kết hợp của lĩnh vực Khai phá dữ liệu và Xử lý ngôn ngữ tự nhiên. Tính đến hiện nay, việc xử lý bài toán phân loại cảm xúc người dùng có nhiều hướng tiếp cận khác nhau và nhiều mức xử lý khác nhau, tuy nhiên các tài liệu liên quan đến việc xử lý đa phần là áp dụng cho tiếng Anh. Việc phân loại cảm xúc người dùng cho Tiếng việt còn nhiều hạn chế.  
Mục đích của luận văn “Phân tích cảm xúc khách hàng qua các bình luận trên trang web bán hàng” là   
Đối tượng nghiên cứu của luận văn là tập trung khai phá phân loại cảm xúc của người dùng khi mua hàng trên các bình luận vào sản phẩm của lazada.vn. Dữ liệu thu thập trên

1. Chương 1: Tổng quan về bài toán phân tích cảm xúc
   1. Giới thiệu chung về phân tích cảm xúc
      1. Giới thiệu

Phân tích tâm lý và khai phá quan điểm người dùng (sentiment analysis hay opinion mining) là một trong những vấn đề được nghiên cứu sôi động nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên, nghiên cứu về các ý kiến, quan điểm, đánh giá, thái độ và cảm xúc của mọi người về một đối tượng.

Cảm xúc được định nghĩa là phản ứng của con người đối với các sự kiện, hiện tượng (kể cả bên trong hoặc bên ngoài cơ thể) cái mà có một ý nghĩa nào đó đối với con người. Cảm xúc, tình cảm là vấn đề được nhiều nhà khoa học quan tâm, nghiên cứu. Vì thế có nhiều quan điểm khác nhau về số lượng các loại cảm xúc. Căn cứ vào tính chất của cảm xúc có thể chia cảm xúc thành 2 loại: *cảm xúc tích cực* và *cảm xúc tiêu cực*. Căn cứ vào biểu hiện và nội dung, chúng ta có thể chia cảm xúc thành 6 loại cơ bản: *vui, buồn, giận dữ, ngạc nhiên, ghét, sợ hãi.* Theo nghiên cứu của W. Gerrod Parrot [W. Gerrod Parrot. “Emotions In Social Psychology”, 2001]. Từ những cảm xúc cơ bản nhưng dưới sự tác động của các kích thích khác nhau trong những điều kiện, hoàn cảnh khác nhau mà cảm xúc của con người cũng có lúc đan xen, pha lẫn nhiều cảm xúc khác loại nhưng cùng tồn tại trong một thời điểm. Và chính điều này đã tạo ra hàng loạt các cảm xúc khác.

Các đối tượng ở đây có thể là các cá nhân, các sự việc, sự vật, các dịch vụ, sản phẩm, các công ty, tổ chức, hoặc một chủ đề bất kỳ. Hai thuật ngữ Opinion Mining (OM) và Sentiment Analysis (SA) có thể được sử dụng thay thế cho nhau trong các ngữ cảnh sử dụng. Tuy nhiên, một số nhà nghiên cứu cho rằng OM và SA có một điểm khác nhau nhỏ [Tsytsarau, M. and Palpanas, T., Survey on Mining Subjective Data on the Web, Data Mining and Knowledge Discovery, 2012, pp.478-514.]. Trong khi OM trích xuất và phân tích các ý kiến của về một đối tượng thì SA cần phải xác định các ý kiến từ một văn bản trước khi tiến hành trích xuất và phân tích chúng.

Phân tích quan điểm là một lĩnh vực thu hút được sự quan tâm lớn của cộng đồng nghiên cứu nói chung và cộng đồng xử lý ngôn ngữ nói riêng bởi ba yếu tố chính sau:

Thứ nhất, đó là sự đa dạng trong ứng dụng của nó vào nhiều lĩnh vực. Ví dụ như trong kinh doanh, việc phân tích và nắm được các ý kiến, quan điểm của khách hàng có thể giúp các công ty, tập đoàn xây dựng được những sản phẩm chất lượng cao, đáp ứng được nhu cầu của người dùng, có thể đưa ra những giải pháp kịp thời cho các vấn đề liên quan đến sản phẩm, chăm sóc khách hàng,.. trước khi mọi chuyện diễn biến xấu đi vượt tầm kiểm soát hoặc thậm chí giúp dự đoán sản lượng bán hàng trong thời gian tới.

Thứ hai, đó là sự bùng nổ của thông tin và mạng xã hội. Trong lịch sử loài người, đây là thời điểm mà lượng thông tin, lượng quan điểm trên mạng internet đang ngày càng trở nên khổng lồ, cung cấp những dữ liệu phong phú, thời gian thực, đa dạng mà không có nó, việc nghiên cứu, phân tích quan điểm người dùng là vô nghĩa. Cộng đồng người dùng Internet ngày càng phát triển và hoạt động tích cực trên các kênh mạng xã hội như Facebook, Twitter, forums, các trang báo,.. với rất nhiều các ý kiến, quan điểm riêng về mọi vấn đề trong xã hội. Tuy nhiên, vấn đề đặt ra là: mặc dù kho dữ liệu khổng lồ này (big data) chứa rất nhiều thông tin, bên cạnh các thông tin hữu ích, được các cá nhân, công ty, tổ chức quan tâm là các thông tin rác không cần thiết. Bài toán đặt ra là làm sao có thể lọc được các thông tin hữu ích này từ kho dữ liệu khổng lồ đó.

Thứ ba, đó là sự thách thức của bài toán. Phân tích quan điểm người dùng có thể được chia ra làm nhiều bài toán nhỏ hơn và cũng đầy thách thức với các nhà nghiên cứu như các bài toán phân lớp chủ quan và khách quan (subjectivity classification), phân lớp ý kiến trái chiều (sentiment polarity classification), phát hiện ý kiến rác (spam opinion detection) [*Rădulescu, C., Dinsoreanu, M. and Potolea, R. (2014), “Identification of Spam Comments Using Natural Language Processing Techniques”, Intelligent Computer Communication and Processing (ICCP), pp. 29-35.*], tóm tắt và tổng hợp quan điểm (opinion summarization), phân tích tính đa diện của của một ý kiến (dual sentiment analysis)[ *Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y. and Li, T. (2015), “Dual Sentiment Analysis: Considering Two Sides of One Review.”, IEEE transactions on knowledge and data engineering, 27(8), pp.2120-2133.*],…

Quan điểm được chia làm hai loại: tích cực (positive) và tiêu cực (negative).

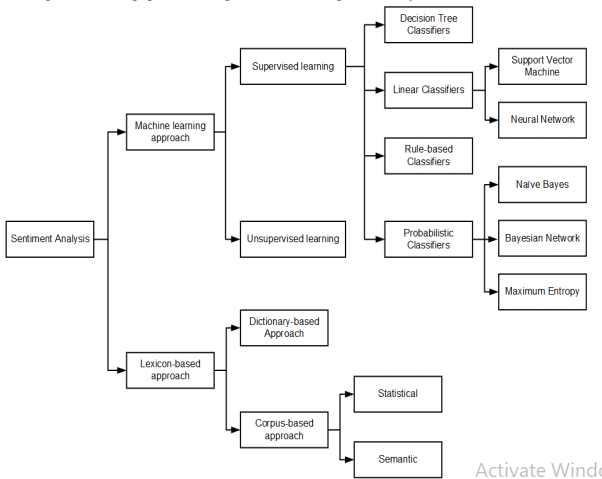
Ngoài hai trạng thái này, một câu hoặc văn bản được xếp vào dạng trung lập (neutral).

Bài toán phân tích quan điểm người dùng thường được tiếp cận và giải quyết ở ba mức độ [5]:

* **Mức độ văn bản, tài liệu** (Document level): ở mức độ này, bài toán cần phân loại xem một văn bản hay tài liệu thể hiện ý kiến tiêu cực hay tích cực. Ví dụ như một bài viết phân tích, đánh giá về kỳ thi đánh giá năng lực do Đại học Quốc gia tổ chức thể hiện ý kiến, nhận định chủ yếu là tốt hay không tốt, tích cực hay tiêu cực. Mức độ này được thực hiện với giả sử rằng tài liệu chỉ đưa ra các quan điểm, ý kiến về một thực thể duy nhất chứ không có sự so sánh giữa các thực thể khác nhau.
* **Mức độ câu** (Sentence level) [3]: các phương pháp được áp dụng cho mức độ tài liệu cũng có thể được áp dụng ở mức độ câu. Trong trường hợp đơn giản, các câu chỉ chứa một ý kiến, quan điểm về một thực thể. Trong các trường hợp phức tạp hơn, một câu có thể có nhiều quan điểm, đánh giá về các khía cạnh khác nhau của một đối tượng hoặc thậm chí có thể có sự thay đổi về quan điểm trong cùng một câu (polarity shifting) [16]. Mức độ phân tích quan điểm cho câu rất gần với bài toán phân lớp chủ quan và khách quan, trong đó chúng ta cần phân loại xem một câu đã cho là chủ quan (có quan điểm, ý kiến riêng) hay khách quan (câu chỉ đưa ra thông tin). Tuy nhiên, các câu khách quan cũng có thể từ đó suy ra quan điểm. Ví dụ như câu: Cơ sở hạ tầng của trường vừa được xây mới một năm nay đã trở nên xập xệ, tồi tàn. Trong câu nói này, cả hai mệnh đề đều là sự việc khách quan trong thực tế nhưng từ đó có thể suy luận ra ý kiến chê bai chất lượng cơ sở vật chất và cũng như cách quản lý chưa sát sao của nhà trường.
* **Mức độ khía cạnh** (Aspect level): nếu với hai mức độ nêu trên, vấn đề được tiếp cận theo hướng kiến trúc của văn bản, ngôn ngữ (câu, đoạn, tài liệu, cú pháp), thì ở mức độ khía cạnh, bài toán tập trung vào chính quan điểm, ý kiến được đưa ra, phân tích ở mức độ sâu hơn, đó là phân tích xem ý kiến tiêu cực hay tích cực của là về chủ đề, đối tượng nào [4].
  1. Các hướng tiếp cận và giải quyết bài toán
     1. Tổng quan

Trong những năm gần đây, có rất nhiều bài báo và các công trình nghiên cứu cải tiến các thuật toán phân tích quan điểm người dùng. Các kỹ thuật này có thể được phân loại như trong Hình 1.1 [7]. Trong đó ta thấy, có hai hướng tiếp cận chính trong các kỹ thuật ứng dụng trong giải quyết bài toán phân lớp quan điểm người dùng, đó là: sử dụng các thuật toán học máy hoặc tiếp cận theo hướng sử dụng các kiến thức về từ vựng và ngữ nghĩa. Trong các thuật toán học máy lại có thể được chia ra thành các thuật toán học có giám sát hay học không giám sát. Ngoài ra, trong một, hai năm trở lại đây bắt đầu xuất hiện các ứng dụng thành công của deep learning vào trong bài toán phân tích quan điểm [12,13] đạt kết quả cao.

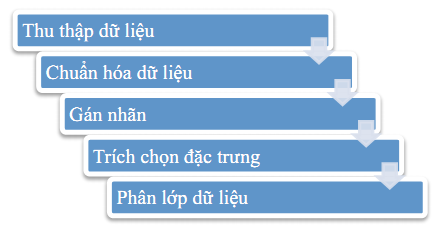
Các thuật toán học máy có giám sát phổ biến được sử dụng trong giải quyết bài toán phân lớp quan điểm là: Naïve Bayes, Maximum Entropy, Support Vector Machine (SVM) [9]. Các thuật toán này được đánh giá cao về tính chính xác và hiệu quả trong giải quyết bài toán phân lớp quan điểm người dùng. Trong mục này, chúng tôi sẽ giới thiệu tổng quan về các giải thuật học có giám sát này.



**Các kỹ thuật sử dụng trong giải quyết bài toán phân lớp quan điểm**

* + 1. Mô hình xử lý dữ liệu cho bài toán phân loại cảm xúc

Mặc dù có những hướng tiếp cận để giải quyết bài toán khác nhau, tuy nhiên nhóm nghiên cứu của Pravesh đã đưa ra mô hình chung cho bài toán này trong hội nghị quốc tế về công nghệ tính toán mềm năm 2014 (International Journal on Soft Computing-IJSC) [18=Singh, Pravesh Kumar, and Mohd Shahid Husain. "METHODOLOGICAL STUDY OF OPINION MINING AND SENTIMENT ANALYSIS TECHNIQUES."International Journal on Soft Computing 5.1, 2014.]



Mô hình xử lý dữ liệu cho bài toán phân loại cảm xúc

Ý nghĩa và nhiệm vụ từng tiến trình cụ thể như sau:

* + - 1. Thu thập dữ liệu

Bước này sẽ thu thập các trang web chứa bình luận người dùng. Để thực hiện bước này có thể sử dụng cơ sở dữ liệu của trang web hoặc sử dụng Crawler hoặc Agent để thực hiện tự động. Kết quả bước này sẽ thu thập được toàn bộ các định dạng dữ liệu của website đối tượng ở dạng HTML hay TXT, dữ liệu này sẽ được đi qua xử lý ở các bước sau.

* + - 1. Chuẩn hóa dữ liệu

Về mặt cấu trúc, dữ liệu thu được ở bước thứ nhất ở định dạng HTML hoặc PHP, do đó bước Chuẩn hóa dữ liệu sẽ lọc lấy những thông tin cần thiết như thời gian, tiêu đề bài viết cùng với bình luận người dùng. Ngoài ra, những bình luận người dùng thu thập được có thể không thỏa mãn những yêu cầu về ngữ pháp hoặc ngữ nghĩa. Bước này cũng sẽ loại bỏ những mẫu bình luận không phù hợp hoặc sửa đổi (như thêm dấu với Tiếng Việt) để đảm bảo dữ liệu thu thập được phù hợp với việc Gán nhãn hay Trích chọn đặc trưng ở bước sau.

* + - 1. Gán nhãn dữ liệu

Với cách tiếp cận bằng bộ từ vựng, bước này sẽ sử dụng dữ liệu từ bộ từ vựng đó để so khớp hay gán nhãn những từ trong văn bản tương ứng trong từ điển. Với phương pháp tiếp cận bằng thuật toán học máy, dữ liệu sẽ được gán nhãn dựa vào đặc trưng ngôn ngữ của văn bản. Cụ thể như gán nhãn câu, gán nhãn từ và từ loại hay các biểu tượng cảm xúc trong bình luận.

* + - 1. Trích chọn đặc trưng

Dựa trên dữ liệu được gán nhãn, đặc trưng để phân lớp là trọng số hướng ngữ nghĩa của bình luận với cách sử dụng bộ từ vựng. Các thuật toán học máy sẽ sử dụng đặc trưng ngôn ngữ như n-grams sau khi tách từ hay gán nhãn từ loại để huấn luyện và kiểm thử với bộ dữ liệu chuẩn được chọn ra từ các bước trước đó.

* + - 1. Phân lớp dữ liệu

Dữ liệu đầu vào đã qua các bước tiền xử lý sẽ qua bộ phân lớp sử dụng các thuật toán học máy. Với cách tiếp cận bằng bộ từ vựng, bước này sẽ áp dụng các luật ngữ pháp hay các quy tắc thay đổi ngữ nghĩa để tính ra trọng số cuối cùng sau đó quyết định hướng ngữ nghĩa của bình luận.

* 1. Mô hình phân lớp Naïve Bayes

Bộ phân lớp quan điểm Naïve Bayes được xây dựng dựa trên lý thuyết Bayes về xác suất có điều kiện và sử dụng mô hình “bag of words” để phân loại văn bản:

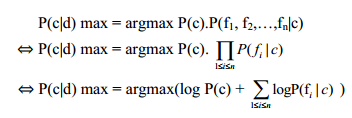


Mục tiêu là tìm được phân lớp c\* sao cho P(c\*|d) là lớn nhất hay xác suất của tài liệu d thuộc lớp c\* là lớn nhất.

Từ công thức trên ta có thể nhận thấy P(d) không đóng vai trò gì trong việc quyết định

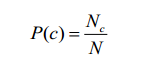
phân lớp c -> P(c|d) lớn nhất ⟺ P(c).P(d|c) lớn nhất.

Để có thể xấp xỉ giá trị của P(d|c), thuật toán Naïve Bayes giả sử rằng: các vector đặc trưng fi của một tài liệu khi đã biết phân lớp là độc lập với nhau. Từ đó ta có công thức:

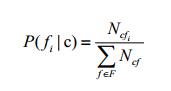


Trong đó f là các vector đặc trưng cho tài liệu d.

Khi tiến hành huấn luyện, thuật toán sử dụng phương pháp xấp xỉ hợp lý cực đại MLE (Maximum Likelihood Estimation) để xấp xỉ P(c) và P(fi|c) cùng thuật toán làm mịn add-one (add-one smoothing). Ta có:



Trong đó Nc là số văn bản được phân loại vào lớp c; N là tổng số văn bản trong tập huấn luyện.



Trong đó Ncfi là số lần xuất hiện của vector đặc trưng i trong tài liệu thuộc phân lớp c.

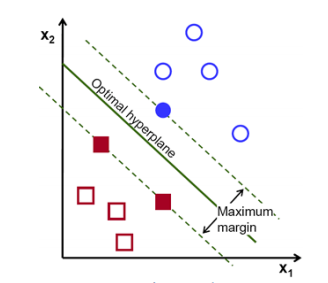
Đánh giá bộ phân lớp sử dụng thuật toán học máy Naive Bayes, ta nhận thấy phương pháp này các ưu điểm như: đơn giản, dễ cài đặt, bộ phân lớp chạy nhanh và cần ít bộ nhớ lưu trữ. Bộ phân lớp cũng không cần nhiều dữ liệu huấn luyện để xấp xỉ được bộ tham số. Tuy nhiên, bộ phân lớp này có nhược điểm là thiếu chính xác do giả thiết độc lập của các vector đặc trưng khi đã biết phân lớp là không có thực trong thực tế.

* 1. Mô hình phân lớp SVM
     1. Giới thiệu về SVM

Máy vector hỗ trợ (Support Vector Machine – SVM) là một phương pháp học máy nổi tiếng được sử dụng để giải quyết bài toán phân lớp, thuật toán được Vladimir N. Vapnik tìm ra và thuật toán SVM tiêu chuẩ hiện nay sử dụng được tìm ra bởi Vapnik và Corinna Cortes vào năm 1995. SVM xuất phát từ lý thuyết học thống kê, dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (Structural Risk Minimisation) và cố gắng phân loại dữ liệu sao cho lỗi xảy ra trên tập kiểm tra là nhỏ nhất (Test Error Minimisation). Nhiều bài toán trong đời sống thực được SVM giải quyết khá thành công như nhận dạng văn bản, hình ảnh, chữ viết tay, phân loại thư rác điện tử, virus…

Thuật toán SVM ban đầu chỉ được thiết kế để giải quyết bài toán phân lớp nhị phân, tức là số lớp hạn chế là hai lớp, với ý tưởng chính như sau:

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector với mỗi điểm là biểu diễn của một dữ liệu, SVM sẽ tìm ra một siêu phẳng f quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt, tương ứng là lớp “+” và lớp “-”. Chất lượng của siêu phẳng được đánh giá bởi khoảng cách lề (margin) giữa hai lớp: khoảng cách càng lớn thì siêu phẳng quyết định càng tốt và chất lượng phân lớp càng cao.



Ví dụ về siêu phẳng trong SVM

Trong ví dụ như Hình 1.2, đường thẳng liền nét tô màu xanh lá chính là siêu phẳng tốt nhất để phân tách dữ liệu thành hai lớp khác nhau. Hai bên của siêu phẳng là hai lề, chứa các vector hỗ trợ (support vectors) – tức là các điểm dữ liệu gần siêu phẳng nhất.

1.4.2 Bài toán phân lớp nhị phân với SVM

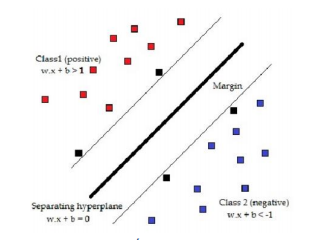
Phát biểu bài toán:

Cho tập mẫu {(x1, y 1), (x2, y2), … (xD, yD)} trong đó xi ∈ RD và yi ∈ {-1, +1}. Giả sử dữ liệu là phân tách tuyến tính, tức là ta có thể phân tách dữ liệu thành hai lớp bằng cách vẽ một đường phẳng trên đồ thị của x1, x2 (với D = 2) hoặc một siêu phẳng trên đồ thị của x1, x2,… xD (với D > 2). Mục đích của thuật toán phân lớp SVM là xây dựng siêu phẳng sao cho khoảng cách lề giữa hai lớp đạt cực đại bằng cách xác định phương trình mô tả siêu phẳng đó trên đồ thị.

Phương pháp giải bài toán:

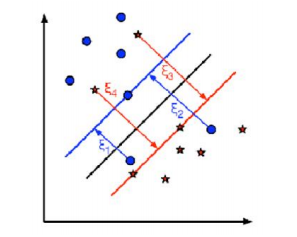
Bài toán này xảy ra ba trường hợp, mỗi trường hợp có một bài toán tối ưu và giải được bài toán này ta sẽ xây dựng được siêu phẳng cần tìm.

* Trường hợp 1: Tập dữ liệu có thể phân chia tuyến tính được mà không có nhiễu, tức là mọi điểm có nhãn “+1” nằm về phía dương trong khi mọi điểm có nhãn “-1” đều nằm về phía âm của mặt phẳng.



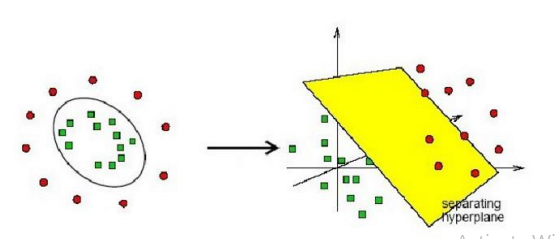
Trường hợp phân chia tuyến tính nhị phân sử dụng SVM không có nhiễu

* Trường hợp 2: Tập dữ liệu có thể phân chia tuyến tính được nhưng có nhiễu, tức là hầu hết các điểm được phân chia đúng bởi siêu phẳng nhưng có tồn tại một vài điểm nhiễu (điểm có nhãn “+1” lại nằm về phía âm của siêu phẳng, hoặc ngược lại).



Trường hợp phân chia tuyến tính nhị phân sử dụng SVM có nhiễu

* Trường hợp 3: Tập dữ liệu không thể phân chia tuyến tính được. Ta sẽ thực hiện phép ánh xạ các vector dữ liệu x vào một không gian khác có nhiều chiều hơn chiều không gian hiện tại sao cho trong không gian này, tập dữ liệu có thể phân chia tuyến tính được.



Trường hợp không thể phân chia tuyến tính nhị phân sử dụng SVM

1.4.3 Bài toán phân lớp đa lớp với SVM

Đối với bài toán phân lớp với số lớp nhiều hơn hai lớp, ta sử dụng kỹ thuật phân đa lớp dạng Multiple Binary Classification với hai chiến lược chính là One-vsOne và One-vs-Rest.

* Chiến thuật One-vs-One: Nếu k là số lớp cần phân tách, chiến lược này sẽ thực hiện k.(k-1)/2 lần phân lớp nhị phân SVM. Cụ thể: ta sẽ bắt cặp từng hai lớp một và sử dụng phương pháp chọn đa số để kết hợp các bộ phân lớp lại thành kết quả phân lớp cuối cùng.
* Chiến thuật One-vs-Rest: Chiến lược này sử dụng (k-1) bộ phân lớp đối với k lớp, tức là chuyển bài toán phân lớp k lớp thành k bài toán phân lớp nhị phân. Trong đó bộ phân lớp nhị phân thứ i được xây dựng trên lớp thứ i và các lớp còn lại.

1.4.4 Đánh giá bộ phân lớp SVM

Bộ phân lớp SVM có các ưu điểm như:

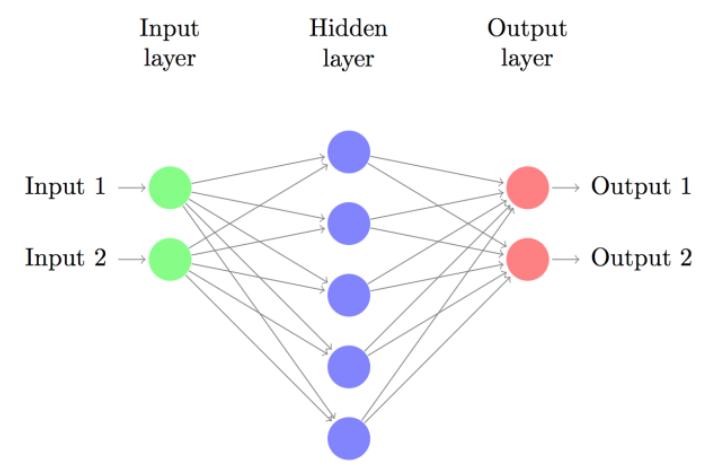
* Độ chính xác phân lớp cao, yêu cầu kích thước bộ dữ liệu huấn luyện nhỏ, dễ áp dụng cho nhiều bài toán.
* Hiệu quả với các bài toán phân lớp dữ liệu có số chiều lớn.
* Hiệu quả với các trường hợp số chiều dữ liệu lớn hơn số lượng mẫu.

Tuy nhiên, bộ phân lớp SVM còn có một số nhược điểm:

* Thời gian huấn luyện lâu, không gian bộ nhớ sử dụng lớn, được thiết kế cho phân lớp nhị phân (trong khi thực tế chủ yếu là phân loại đa lớp).
* Có thể bị overfit trên dữ liệu huấn luyện, nhạy cảm với nhiễu.
  1. Mô hình Neural Network
     1. Kiến trúc mạng nơ-ron nhân tạo

Mạng nơ ron nhân tạo (Artificial Neural Network – ANN) là một mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các Nơ-ron được gắn kết để xử lý thông tin. ANN hoạt động giống như bộ não của con người, được học bởi kinh nghiệm (thông qua việc huấn luyện), có khả năng lưu giữ các tri thức và sử dụng các tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data).

Một mạng nơ-ron là một nhóm các nút nối với nhau, mô phỏng mạng nơ ron thần kinh của não người. Mạng nơ ron nhân tạo được thể hiện thông qua ba thành phần cơ bản: mô hình của nơ ron, cấu trúc và sự liên kết giữa các nơ ron.Trong nhiều trường hợp, mạng nơ ron nhân tạo là một hệ thống thích ứng, tự thay đổi cấu trúc của mình dựa trên các thông tin bên ngoài hay bên trong chạy qua mạng trong quá trình học.



Kiến trúc mạng nơ-ron nhân tạo

Kiến trúc chung của một ANN gồm 3 thành phần đó là **Input Layer**, **Hidden Layer** và **Output Layer** (Hình 2.1)

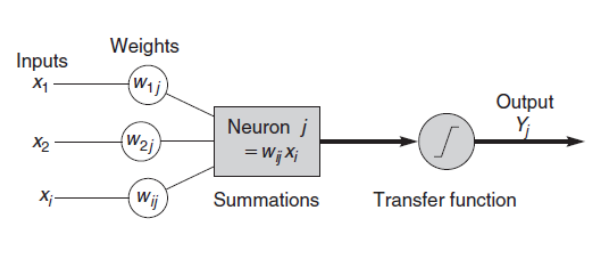
Trong đó, lớp ẩn (**Hidden Layer**) gồm các nơ-ron, nhận dữ liệu input từ các Nơ-ron ở lớp (Layer) trước đó và chuyển đổi các input này cho các lớp xử lý tiếp theo. Trong một mạng ANN có thể có nhiều Hidden Layer.

Lợi thế lớn nhất của các mạng ANN là khả năng được sử dụng như một cơ chế xấp xỉ hàm tùy ý mà “học” được từ các dữ liệu quan sát. Tuy nhiên, sử dụng chúng không đơn giản như vậy, một số các đặc tính và kinh nghiệm khi thiết kế một mạng nơ-ron ANN.

* **Chọn mô hình**: Điều này phụ thuộc vào cách trình bày dữ liệu và các ứng dụng. Mô hình quá phức tạp có xu hướng dẫn đền những thách thức trong quá trình học.
* **Cấu trúc và sự liên kết giữa các nơ-ron**
* **Thuật toán học:** Có hai vấn đề cần học đối với mỗi mạng ANN, đó là học tham số của mô hình (parameter learning) và học cấu trúc (structure learning). Học tham số là thay đổi trọng số của các liên kết giữa các nơ-ron trong một mạng, còn học cấu trúc là việc điều chỉnh cấu trúc mạng bằng việc thay đổi số lớp ẩn, số nơ-ron mỗi lớp và cách liên kết giữa chúng. Hai vấn đề này có thể được thực hiện đồng thời hoặc tách biệt.

Nếu các mô hình, hàm chi phí và thuật toán học được lựa chọn một cách thích hợp, thì mạng ANN sẽ cho kết quả có thể vô cùng mạnh mẽ và hiệu quả.

* + 1. Hoạt động của mạng nơ-ron nhân tạo



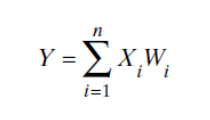
Quá trình xử lý thông tin của một mạng nơ-ron nhân tạo

**Inputs**: Mỗi Input tương ứng với 1 đặc trưng của dữ liệu. Ví dụ như trong ứng dụng của ngân hàng xem xét có chấp nhận cho khách hàng vay tiền hay không thì mỗi input là một thuộc tính của khách hàng như thu nhập, nghề nghiệp, tuổi, số con,…

**Output**: Kết quả của một ANN là một giải pháp cho một vấn đề, ví dụ như với bài toán xem xét chấp nhận cho khách hàng vay tiền hay không thì output là yes hoặc no.

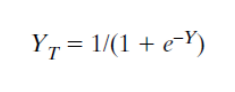
**Connection Weights** (Trọng số liên kết) : Đây là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng, độ mạnh của dữ liệu đầu vào đối với quá trình xử lý thông tin chuyển đổi dữ liệu từ Layer này sang layer khác. Quá trình học của ANN thực ra là quá trình điều chỉnh các trọng số Weight của các dữ liệu đầu vào để có được kết quả mong muốn.

**Summation Function** (Hàm tổng): Tính tổng trọng số của tất cả các input được đưa vào mỗi Nơ-ron. Hàm tổng của một Nơ-ron đối với n input được tính theo công thức sau:



**Transfer Function** (Hàm chuyển đổi): Hàm tổng của một nơ-ron cho biết khả năng kích hoạt của nơ-ron đó còn gọi là kích hoạt bên trong. Các nơ-ron này có thể sinh ra một output hoặc không trong ANN, nói cách khác rằng có thể output của 1 Nơ-ron có thể được chuyển đến layer tiếp trong mạng Nơ-ron theo hoặc không. Mối quan hệ giữa hàm tổng và kết quả output được thể hiện bằng hàm chuyển đổi.

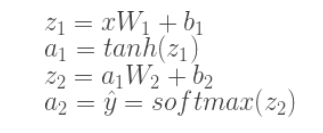
Việc lựa chọn **Transfer Function** có tác động lớn đến kết quả của ANN. Hàm chuyển đổi phi tuyến được sử dụng phổ biến trong ANN là sigmoid (logical activation) function.



Kết quả của Sigmoid Function thuộc khoảng [0,1] nên còn gọi là hàm chuẩn hóa (Normalized Function).

Kết quả xử lý tại các nơ-ron (Output) đôi khi rất lớn, vì vậy transfer function được sử dụng để xử lý output này trước khi chuyển đến layer tiếp theo. Đôi khi thay vì sử dụng Transfer Function người ta sử dụng giá trị ngưỡng (Threshold value) để kiểm soát các output của các neuron tại một layer nào đó trước khi chuyển các output này đến các Layer tiếp theo. Nếu output của một neuron nào đó nhỏ hơn Threshold thì nó sẻ không được chuyển đến Layer tiếp theo.

Mạng nơ-ron của chúng ta dự đoán dựa trên forward propagation là các phép nhân ma trận cùng với activation function để thu được kết quả đầu ra. Nếu input x là vector 2 chiều thì ta có thể tính kết quả dự đoán bằng công thức sau:



Trong đó, zi là input của layer thứ i, ai là output của layer thứ sau khi áp dụng activation function. W1, b1, W2, b2 là các thông số (parameters) cần tìm của mô hình mạng nơ-ron.

Huấn luyện để tìm các thông số cho mô hình tương đương với việc tìm các thông số W1, b1, W2, b2, sao cho độ lỗi của mô hình đạt được là thấp nhất. Ta gọi hàm độ lỗi của mô hình là loss function. Đối với softmax function, ta dùng **cross-entropy loss** (còn gọi là negative log likelihood).

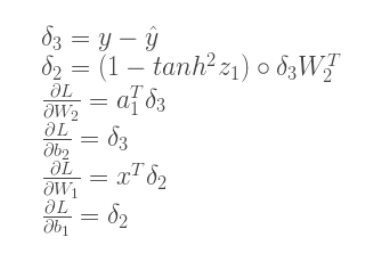
Nếu ta có N dòng dữ liệu huấn luyện, và C nhóm phân lớp (trường hợp này là hai lớp nam, nữ), khi đó loss function giữa giá trị dự đoán và được tính như sau:



Ý nghĩa công thức trên nghĩa là: lấy tổng trên toàn bộ tập huấn luyện và cộng dồn vào hàm loss nếu kết quả phân lớp sai. Độ dị biệt giữa hai giá trị và càng lớn thì độ lỗi càng cao. Mục tiêu của chúng ta là tối thiểu hóa hàm lỗi này. Ta có thể sử dụng phương pháp gradient descent để tối tiểu hóa hàm lỗi. Có hai loại gradient descent, một loại với fixed learning rate được gọi là batch gradient descent, loại còn lại có learning rate thay đổi theo quá trình huấn luyện được gọi là SGD (stochastic gradient descent) hay minibatch gradient descent.

Gradient descent cần các gradient là các vector có được bằng cách lấy đạo hàm của loss function theo từng thông số  Để tính các gradient này, ta sử dụng thuật toán *backpropagation (lan truyền ngược)*. Đây là cách hiệu quả để tính gradient khởi điểm từ output layer.

Áp dụng backpropagation ta có các đại lượng:



* 1. Đánh giá và kết luận

1. Chương 2: Xây dựng hệ thống phân tích cảm xúc dựa trên Neural Network
   1. Mô hình hệ thống
   2. Lấy dữ liệu, tách đặc trưng
   3. Huấn luyện
   4. Phân loại
   5. Kết luận và đánh giá
2. Chương 3: Cài đặt hệ thống
   1. Môi trường cài đặt
   2. Django
   3. SQLite
   4. Cài đặt
   5. Kết luận và đánh giá
3. Chương 4: Đánh giá
   1. Hoạt động của hệ thống
   2. Đánh giá hiệu quả
   3. Kết luận
4. Chương 5: Kết luận chung
5. Chương 6: Tài liệu tham khảo

3. Devi, G.D. and Rasheed, A.A. (2015), “A Survey on Sentiment Analysis and

Opinion Mining.”, International journal for research in emerging science and technology 2(8), pp. 26-31.

4. Hu, M. and Liu, B. (2004), “Mining and Summarizing Customer Reviews.”, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177.

5. Liu, B. (2012), Sentiment Analysis and Opinion Mining, Morgan & Claypool.

6. Ben Krose and Patrick van der Smagt, An Introduction to Neural Networks, @ 1996 University of Amsterdam.

7. Medhat, W., Hassan, A. and Korashy, H. (2014), “Sentiment Analysis Algorithms and Applications: A Survey.”, Ain Shams Engineering Journal 5(4), pp. 1093-1113.

8. Robert J. Schallkoff – Artificial Neural Networks, the McGraw – Hill

Companies, Inc 1997.

12. Severyn, A. and Moschitti, A. (2015), “Twitter Sentiment Analysis with Deep Convolutional Neural Networks.”, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 959-962.

13. Tang, D., Qin, B. and Liu, T. (2015), “Deep Learning for Sentiment

Analysis: Successful Approaches and Future Challenges.”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5(6), pp. 292-303.

16. Zhang, X., Li, S., Zhou, G. and Zhao, H. (2011), “Polarity Shifting: Corpus

Construction and Analysis.”, Asian Language Processing (IALP), pp. 272-275.