# ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

A blue logo with a planet and a flower

Description automatically generated**---------□&□---------**

**ĐỒ ÁN MÔN HỌC MẠNG XÃ HỘI**

**ĐỀ TÀI NGHIÊN CỨU**

**Sentiment Analysis of Twitter Data for Customer Service Applications**

*Giảng viên hướng dẫn:* **ThS. Thái Bảo Trân**

*Nhóm sinh viên thực hiện:* **Nhóm 12 - IS353.O21**

**Nguyễn Văn Thịnh** 19522286

**Đinh Công Vĩnh An** 21521803

*Thành phố Hồ Chí Minh, tháng 6 năm 2024*

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

# LỜI MỞ ĐẦU

Trong thời đại kỹ thuật số ngày nay, mạng xã hội đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của con người. Trong số đó, Twitter nổi bật như một nền tảng giao tiếp nhanh chóng và hiệu quả, nơi người dùng có thể chia sẻ ý kiến, cảm xúc và phản hồi về mọi vấn đề từ cuộc sống cá nhân đến các dịch vụ khách hàng. Việc phân tích cảm xúc[**Sentiment Analysis]** từ dữ liệu Twitter không chỉ cung cấp những thông tin quý giá về thái độ và tâm trạng của người dùng mà còn mở ra những cơ hội lớn cho các doanh nghiệp trong việc cải thiện dịch vụ khách hàng.

Lời đầu tiên chúng em xin chân thành gửi lời cảm ơn đến với giảng viên môn học là cô Thái Bảo Trân. Cảm ơn cô đã giảng dạy, truyền đạt rất nhiều kiến thức bổ ích về mạng xã hội, những kiến thức này vô cùng cần thiết cho chúng em thực hiện đồ án. Cảm ơn cô đã hướng dẫn nhiệt tình, giải đáp thắc mắc cũng như hỗ trợ chúng em rất nhiều trong quá trình thực hiện đồ án.

Với khả năng và thời gian có hạn nên không thể tránh khỏi những thiếu sót, chúng em rất mong được sự quan tâm, giúp đỡ và thông cảm của cô để chúng em hoàn thiện hơn về đồ án của nhóm mình. Một lần nữa, chúng em xin chân thành cảm ơn cô!

Thành phố Hồ Chí Minh, tháng 6 năm 2024

Nhóm sinh viên thực hiện

Nhóm 12

# BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ tên | Phân công | Đánh giá |
| 19522286 | Nguyễn Văn Thịnh | Tìm hiểu bài báo khoa học và thuật toán khám phá cộng đồng  Phân tích đề tài nghiên cứu và cách triển khai mô phỏng thuật toán  Tìm tài liệu thực nghiệm cho mô phỏng thuật toán  Thực hiện phần tài liệu về phương pháp nghiên cứu, các quy trình, chỉ số thuật toán và các thuộc tính trong khám phá cộng đồng  Thực hiện tài liệu báo cáo , nội dung thuyết trình | Hoàn thành 100% |
| 21521803 | Đinh Công Vĩnh An | Tìm hiểu bài báo khoa học và thuật toán khám phá cộng đồng  Phân tích đề tài nghiên cứu và cách triển khai mô phỏng thuật toán  Tìm tài liệu thực nghiệm cho mô phỏng thuật toán  Thực hiện phần tài liệu về phương pháp nghiên cứu, các quy trình, chỉ số thuật toán và các thuộc tính trong khám phá cộng đồng  Thực hiện tài liệu báo cáo , nội dung thuyết trình | Hoàn thành 100% |

# MỤC LỤC

[ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH 1](#_Toc168435077)

[LỜI MỞ ĐẦU 3](#_Toc168435078)

[BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN 4](#_Toc168435079)

[MỤC LỤC 5](#_Toc168435080)

[CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 7](#_Toc168435081)

[1.1 Thông tin tác giả 7](#_Toc168435082)

[1.2 Thông tin bài báo 7](#_Toc168435083)

[1.3 Tổng quan nội dung 7](#_Toc168435084)

[CHƯƠNG 2: NỘI DUNG NGHIÊN CỨU 8](#_Toc168435085)

[2.1 Các thuật ngữ liên quan 8](#_Toc168435086)

[2.2 Mục tiêu nghiên cứu 10](#_Toc168435087)

[2.3 Phương pháp nghiên cứu 11](#_Toc168435088)

[CHƯƠNG 3: CHUẨN BỊ DATASET 13](#_Toc168435089)

[3.1 Mô tả Dataset 13](#_Toc168435090)

[3.2 Xử lý dữ liệu 14](#_Toc168435091)

[CHƯƠNG 4: SENTIMENT ANALYSIS MODEL 16](#_Toc168435092)

[4.1 Support Vector Machine (SVM) 16](#_Toc168435093)

[4.2 Convolutional Neural Network (CNN) 19](#_Toc168435094)

[4.2.1 Kiến trúc CNN 19](#_Toc168435095)

[4.2.2 Huấn luyện và đánh giá 20](#_Toc168435096)

[4.3 Recurrent Neural Network (RNN) - Long Short Term Memory (LSTM) 21](#_Toc168435097)

[4.3.1 RNN 22](#_Toc168435098)

[4.3.2 LSTM 25](#_Toc168435099)

[CHƯƠNG 5: TRỰC QUAN HÓA 28](#_Toc168435100)

[5.1 Trực quan hóa dữ liệu Keyword 28](#_Toc168435101)

[5.2 Trực quan hóa bằng đồ thị 29](#_Toc168435102)

[CHƯƠNG 6: HƯỚNG NGHIÊN CỨU 31](#_Toc168435103)

[6.1 Cải thiện mô hình 31](#_Toc168435104)

[6.2 Mở rộng ứng dụng 31](#_Toc168435105)

[CHƯƠNG 6: TÀI LIỆU THAM KHẢO 33](#_Toc168435106)

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

Nhóm thực hiện nghiên cứu đề tài **“****Sentiment Analysis of Twitter Data for Customer Service Applications”** dựa trên bài báo nghiên cứu khoa học **“Sentiment Analysis of Twitter Data”**

Truy cập bài báo: <https://www.mdpi.com/2076-3417/12/22/11775>

## 1.1 Thông tin tác giả

Bài viết được viết bởi Yili Wang, Jiaxuan Guo, Chengsheng Yuan và Baozhu Li, những nhà nghiên cứu hàng đầu trong lĩnh vực mạng xã hội. Họ đến từ Đại học Khoa học và Công nghệ Thông tin Nam Kinh, Đại học Phúc Đán – Trung quốc, Đại học Sungkyunkwan - Hàn Quốc.

## 1.2 Thông tin bài báo

Bài viết có tựa đề “Sentiment Analysis of Twitter Data” được công bố trên tạp chí Multidisciplinary Digital Publishing Institute (MDPI) vào ngày 19 tháng 11 năm 2022. Bài viết là kết quả của một dự án nghiên cứu lâu dài, trong đó các tác giả đã dành rất nhiều thời gian và công sức để phát triển và kiểm tra thuật toán Phân tích cảm xúc (SA) cho mạng xã hội Twitter.

## 1.3 Tổng quan nội dung

Đề tài “Sentiment Analysis of Twitter Data for Customer Service Applications” (Phân Tích Cảm Xúc Dữ Liệu Twitter cho Ứng Dụng Dịch Vụ Khách Hàng) nhằm mục đích nghiên cứu và ứng dụng các kỹ thuật phân tích cảm xúc để đánh giá phản hồi của khách hàng trên Twitter. Thông qua việc áp dụng các phương pháp học máy và xử lý ngôn ngữ tự nhiên (NLP), chúng ta có thể tự động xác định và phân loại các phản hồi tích cực, tiêu cực và trung tính từ người dùng. Điều này không chỉ giúp doanh nghiệp hiểu rõ hơn về cảm nhận của khách hàng mà còn hỗ trợ trong việc cải thiện chất lượng dịch vụ, nâng cao trải nghiệm người dùng và tăng cường sự hài lòng của khách hàng.

Nội dung của đồ án sẽ bao gồm việc thu thập và tiền xử lý dữ liệu Twitter, xây dựng mô hình phân tích cảm xúc, và đánh giá hiệu quả của mô hình thông qua các chỉ số đo lường.

# CHƯƠNG 2: NỘI DUNG NGHIÊN CỨU

Do sự gia tăng bùng nổ gần đây của các Dịch Vụ Mạng Xã Hội (SNS), một lượng lớn dữ liệu do người dùng tạo ra, chẳng hạn như bình luận và đánh giá, đang được tạo ra liên tục. Ý kiến và cảm xúc của mọi người được thể hiện trong thông tin, chủ yếu dựa trên một đối tượng quan tâm chung. Những dữ liệu này đã trở thành kho tàng thông tin, mang lại nhiều cơ hội để phân tích phản ứng của mọi người, đặc biệt là có lợi trong việc dự báo doanh số bán hàng, xu hướng thị trường chứng khoán, và kết quả của các cuộc bầu cử chính trị. Có hơn 300 triệu người dùng Twitter hoạt động, khiến nó trở thành một trong những dịch vụ micro-blogging phổ biến nhất. Do tầm quan trọng của nó trong việc nhận thức suy nghĩ và thái độ của mọi người, Phân Tích Cảm Xúc Dựa Trên Twitter (TSA) đã thu hút rất nhiều sự chú ý.

## 2.1 Các thuật ngữ liên quan

**Phân tích Cảm xúc trên Twitter (Twitter-based Sentiment Analysis - TSA):** TSA đề cập đến quá trình sử dụng các phương pháp tính toán để phân tích và diễn giải cảm xúc thể hiện trong dữ liệu từ Twitter. Nó liên quan đến việc phân loại các tweet thành các cảm xúc tích cực, tiêu cực hoặc trung tính(positive, negative, neutral).

A diagram of a process

Description automatically generated

**Dịch Vụ Mạng Xã Hội (Social Networking Service - SNS):** SNS là các nền tảng trực tuyến cho phép người dùng tạo hồ sơ, chia sẻ nội dung và tương tác với người khác. Twitter là một ví dụ nổi bật về SNS.

**Phân Tích Cảm Xúc (Sentiment Analysis - SA):** Là quá trình phát hiện và trích xuất thông tin chủ quan từ văn bản. Nó nhằm xác định cảm xúc thể hiện trong văn bản, liệu nó là tích cực, tiêu cực hay trung tính. TSA là một ứng dụng cụ thể của SA tập trung vào Twitter.

**Khai Thác Ý Kiến (Opinion Mining - OM):** Thường được sử dụng thay thế cho SA, liên quan đến việc trích xuất và phân tích các ý kiến từ dữ liệu văn bản. Nó đặc biệt tập trung vào việc xác định các phát biểu chủ quan trong dữ liệu. OM rất quan trọng trong việc hiểu phản hồi của người tiêu dùng, ý kiến công chúng và các cuộc thảo luận trên mạng xã hội.

**Xử Lý Ngôn Ngữ Tự Nhiên (Natural Language Processing - NLP):** là một lĩnh vực của trí tuệ nhân tạo tập trung vào tương tác giữa máy tính và ngôn ngữ của con người. Nó bao gồm nhiều kỹ thuật để xử lý và phân tích dữ liệu ngôn ngữ tự nhiên. Các kỹ thuật NLP là cơ sở cho việc thực hiện TSA, vì chúng cho phép xử lý và phân tích tự động dữ liệu văn bản từ Twitter.

**Phương Pháp Dựa Trên Học Máy (Machine Learning-Based Approach):** Phương pháp này bao gồm việc huấn luyện các mô hình học máy trên các tập dữ liệu đã được gán nhãn để dự đoán cảm xúc. Các thuật toán phổ biến bao gồm Naïve Bayes (NB), Support Vector Machine (SVM), và Neural Networks (NN).

**Phương Pháp Dựa Trên Từ Vựng (Lexicon-Based Approach):** Phương pháp này sử dụng các từ điển cảm xúc đã được xác định trước, là tập hợp các từ liên quan đến cảm xúc tích cực hoặc tiêu cực, để xác định cảm xúc của văn bản.

**Phương Pháp Lai (Hybrid Approach):** Các phương pháp lai kết hợp giữa học máy và từ vựng để tận dụng điểm mạnh của cả hai. Những phương pháp này thường mang lại hiệu suất phân loại tốt hơn. Các phương pháp lai giải quyết được hạn chế của các phương pháp đơn lẻ và hiệu quả trong các nhiệm vụ phân tích cảm xúc phức tạp.

**N-gram:** Một N-gram là một chuỗi liên tục của n mục từ một văn bản cho trước. Unigram, bigram, và trigram lần lượt là các chuỗi của một, hai, và ba mục. N-gram được sử dụng trong trích xuất đặc trưng cho phân tích cảm xúc, giúp nắm bắt ngữ cảnh và mối quan hệ giữa các từ.

**Gắn Thẻ Phần Từ Loại (Part of Speech - POS) Tagging:** POS tagging liên quan đến việc gắn các phần từ loại (như danh từ, động từ, tính từ) cho mỗi từ trong văn bản.

**Biểu Diễn Đặc Trưng (Feature Representation – FP):** liên quan đến việc chuyển đổi văn bản thành các đặc trưng số có thể sử dụng để huấn luyện các mô hình học máy. Các kỹ thuật phổ biến bao gồm N-gram và POS tagging. FP rất quan trọng cho hiệu suất của các mô hình phân tích cảm xúc.

**Naïve Bayes (NB):** NB là một bộ phân loại xác suất dựa trên việc áp dụng định lý Bayes với các giả định mạnh mẽ về tính độc lập giữa các đặc trưng. NB được sử dụng phổ biến trong TSA do sự đơn giản và hiệu quả của nó trong các nhiệm vụ phân loại văn bản.

**Support Vector Machine (SVM):** SVM là một mô hình học có giám sát phân tích dữ liệu để phân loại và phân tích hồi quy. Nó hiệu quả trong các không gian có chiều cao. SVM được sử dụng phổ biến trong TSA nhờ tính chính xác và độ bền cao trong việc phân loại cảm xúc.

**Neural Network (NN):** NN là các hệ thống tính toán lấy cảm hứng từ mạng lưới thần kinh sinh học của não động vật. Chúng có khả năng nhận dạng mẫu và học từ dữ liệu. NN, đặc biệt là các mô hình học sâu, ngày càng được sử dụng trong TSA nhờ khả năng nắm bắt các mẫu phức tạp trong dữ liệu văn bản.

## 2.2 Mục tiêu nghiên cứu

Mục tiêu của bài báo là cung cấp một cái nhìn tổng quan gần như toàn diện về các kỹ thuật và lĩnh vực liên quan đến phân tích cảm xúc trên Twitter (Twitter Sentiment Analysis - TSA). Cụ thể, bài báo tập trung vào các điểm sau:

**Khảo sát và Phân loại các Phương pháp TSA:** Bài báo tiến hành khảo sát các nghiên cứu gần đây nhất trong lĩnh vực TSA, phân loại và sắp xếp các thuật toán và ứng dụng mới được đề xuất. Mục đích là phân loại chi tiết các bài báo gần đây và miêu tả hướng nghiên cứu hiện tại trong lĩnh vực TSA​​.

**Giới thiệu Các Khái niệm và Kỹ thuật Cơ bản:** Bài báo cung cấp một giới thiệu ngắn gọn về các khái niệm cơ bản và kỹ thuật trong phân tích cảm xúc (Sentiment Analysis - SA), nhằm giúp người đọc hiểu rõ hơn về TSA và các ứng dụng gần đây​​.

**Đánh Giá Các Phương Pháp Hiện Tại:** Bài báo đánh giá các phương pháp hiện tại, đặc biệt tập trung vào ba phương pháp chính: dựa trên học máy (machine-learning-based), dựa trên từ vựng (lexicon-based), và phương pháp kết hợp (hybrid approaches). Bài báo so sánh và phân tích hiệu suất của các phương pháp này​​.

A diagram of a variety of learning styles

Description automatically generated with medium confidence

**Định Hướng Nghiên Cứu Tương Lai:** Bài báo cũng đưa ra các câu hỏi nghiên cứu và thảo luận về hướng phát triển tương lai của TSA, nhằm định hướng cho các nghiên cứu tiếp theo trong lĩnh vực này​​.

Thông qua các mục tiêu này, bài báo hy vọng sẽ cung cấp một tài liệu tham khảo quan trọng cho các nhà nghiên cứu và thực hành trong lĩnh vực TSA, giúp họ có cái nhìn toàn diện và cập nhật về các kỹ thuật và phương pháp hiện có.

## 2.3 Phương pháp nghiên cứu

Bài báo sử dụng phương pháp nghiên cứu tổng quan và phân loại các phương pháp phân tích cảm xúc trên Twitter (Twitter Sentiment Analysis - TSA). Cụ thể, phương pháp nghiên cứu của bài báo được chia thành các phần sau:

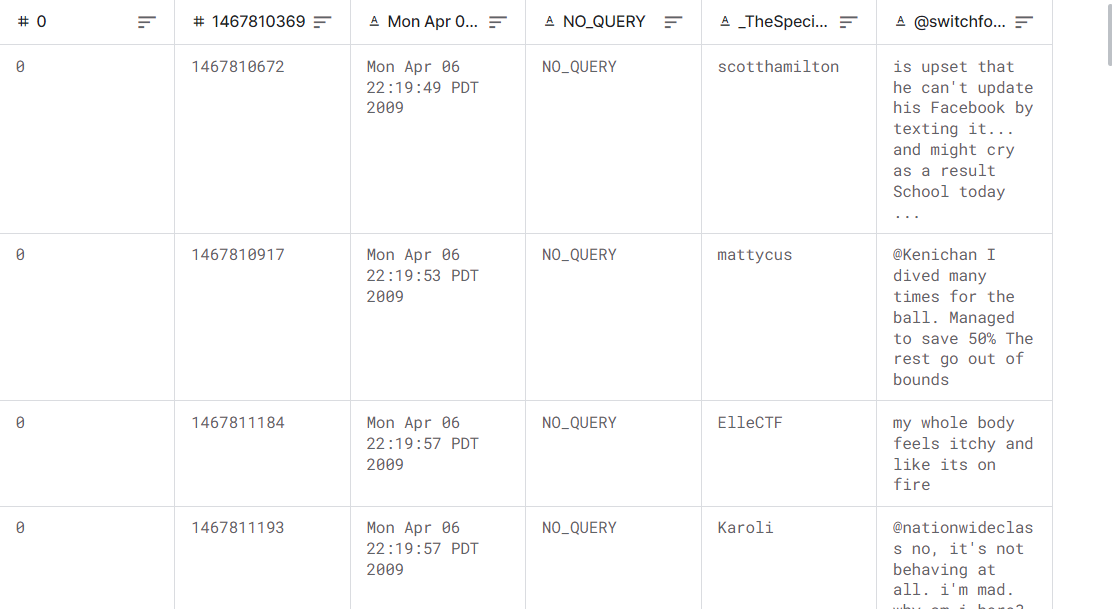
1. **Phương pháp Học Máy (Machine-Learning-Based Approach):**
   * Phương pháp này sử dụng các thuật toán học máy truyền thống được huấn luyện bằng một tập hợp các đặc trưng có sẵn để dự đoán độ phân cực cảm xúc của một đoạn văn bản nhất định. Bài báo chỉ ra rằng hiệu suất của sự kết hợp nhiều bộ phân loại thường tốt hơn so với việc sử dụng một bộ phân loại đơn lẻ. Tuy nhiên, phương pháp này có những hạn chế như cần một lượng lớn dữ liệu được gán nhãn thủ công, và độ chính xác của dự đoán phụ thuộc nhiều vào bộ dữ liệu huấn luyện​​.
2. **Phương pháp Dựa Trên Từ Vựng (Lexicon-Based Approach):**
   * Phương pháp này sử dụng một tập hợp các từ vựng cảm xúc để đo lường cường độ của cảm xúc được biểu đạt. Các từ vựng cảm xúc được xây dựng dựa trên hai kỹ thuật chính: dựa trên từ điển và dựa trên tập hợp văn bản. Ưu điểm của phương pháp này là không cần dữ liệu được gán nhãn, nhưng nhược điểm là hiệu suất có thể bị giảm nếu các từ không có trong từ điển hoặc không xem xét được ngữ cảnh của các từ​​.
3. **Phương pháp Kết Hợp (Hybrid Approach):**
   * Phương pháp kết hợp các kỹ thuật học máy và từ vựng để cải thiện hiệu suất phân loại. Phương pháp này đã được chứng minh là cải thiện hiệu suất trong các lĩnh vực cụ thể nhưng đòi hỏi chi phí tính toán cao​​.
4. **Phương pháp Khác:**
   * Bài báo cũng đề cập đến một số phương pháp khác như phương pháp dựa trên đồ thị (graph-based approaches), xây dựng một đồ thị xã hội kết nối để gắn nhãn hiệu quả, dựa trên giả định rằng con người có ảnh hưởng lẫn nhau​​.

Bằng cách sử dụng các phương pháp này, bài báo tiến hành phân loại và đánh giá các nghiên cứu gần đây nhất trong lĩnh vực TSA, cung cấp một cái nhìn tổng quan và so sánh chi tiết về các kỹ thuật và phương pháp hiện có​​.

# CHƯƠNG 3: CHUẨN BỊ DATASET

Sử dụng bộ Dataset được chuẩn bị sẵn:

[https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv](https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv%20)



## 3.1 Mô tả Dataset

Dataset bao gồm 1.600.000 tweet được thu thập từ Twitter. Mỗi Tweet được biểu diễn dưới dạng một chuỗi văn bản, bao gồm nội dung, tên người dùng được tag.

Dataset bao gồm 6 cột từ trái sang phải:

1. Cảm xúc của tweet, được phân loại từ 0-4 (tiêu cực – trung tính – tích cực).
2. ID của tweet.
3. Thời gian đăng tweet.
4. Truy vấn.
5. Tên người dùng (username).
6. Tweet.

## 3.2 Xử lý dữ liệu

A screenshot of a computer

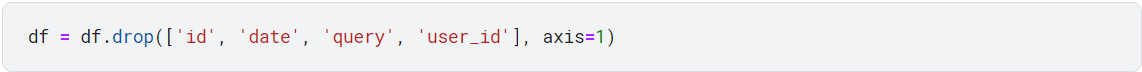
Description automatically generated

Đọc tập dữ liệu phân tích cảm tính từ tệp CSV, xem qua dữ liệu ban đầu, gán tên cột có ý nghĩa để hiểu rõ hơn và sau đó hiển thị lại các hàng ban đầu cùng với nhãn

A screenshot of a computer

Description automatically generated

Tạo danh sách các chỉ mục ngẫu nhiên trong phạm vi độ dài chuỗi



Loại bỏ các cột không cần thiết.

# CHƯƠNG 4: SENTIMENT ANALYSIS MODEL

Chương này trình bày các phương pháp được sử dụng để kiểm tra trên tập dữ liệu tích hợp sẵn của bài báo. Tác giả nhận ra rằng các vấn đề quan tâm của họ được giải quyết theo các vấn đề phân loại. Do đó, họ tiến hành sử dụng học máy để giải quyết các vấn đề phân loại cho bài báo này.

Hiện nay, có hai phương pháp chính trong Machine learning: Supervised learning (học có giám sát) và (Unsupervised learning) học không giám sát. Học có giám sát là một thuật toán dự đoán đầu ra cho dữ liệu mới dựa trên danh sách kẹp đầu vào và đầu ra đã được đào tạo trước đây. Học không giám sát, thuật toán sẽ dự đoán đầu ra dựa trên cấu trúc và đặc điểm của dữ liệu mà không cần biết đầu ra trước. Ngoài ra, một số phương pháp tiếp cận trong phương pháp học máy là học bán giám sát và học củng cố. Bên cạnh đó, trong những năm gần đây, xu hướng các nhà khoa học áp dụng học sâu vào các dự án nghiên cứu đã trở nên phổ biến hơn và có độ chính xác cao hơn so với các thuật toán khác.

## 4.1 Support Vector Machine (SVM)

SVM là một thuật toán học máy giám sát được sử dụng cho cả các vấn đề phân loại và hồi quy. Tuy nhiên, SVM chủ yếu được sử dụng trong các vấn đề phân loại. Hiện nay, SVM được coi là một thuật toán mạnh mẽ trong lĩnh vực Machine learning.

Giả sử chúng ta có hai phân lớp con khác nhau được mô tả bởi các điểm trong không gian đa chiều. Yêu cầu của chúng ta là tìm một mặt phẳng để phân loại chính xác hai phân lớp con đó, tức là mặt phẳng sẽ tách các điểm thuộc cùng một phân lớp nằm cùng một phía của mặt phẳng và hai phân lớp con khác nhau sẽ nằm ở các phía khác nhau của mặt phẳng. Trong hình dưới, chúng ta có thể thấy rằng có nhiều mặt phẳng được tìm thấy để giải quyết các vấn đề trên. A close-up of a line

Description automatically generated



Tuy nhiên, trong những mặt phẳng đó, chúng ta cần tìm mặt phẳng có thể cắt tốt nhất và hiệu quả nhất. Mặt phẳng tốt nhất để tìm kiếm là mặt phẳng phân chia sao cho khoảng cách từ điểm gần nhất của các lớp đến mặt phẳng bằng nhau và khoảng cách đó được gọi là biên độ (margin). Sau đó, chúng ta cần một mặt phẳng mà tại đó biên độ lớn nhất sẽ cho phép xếp lớp tốt hơn. Thuật toán SVM được phát triển từ việc tìm kiếm một siêu mặt phẳng sao cho biên độ là lớn nhất.

A diagram of lines and dots

Description automatically generated

Giả sử ta có: và giá trị của . Theo thuật toán SVM, ta sẽ tìm được 1 siêu mặt phẳng chia cắt 2 lớp con và có biên độ lớn nhất. Phương trình để tìm siêu mặt phẳng đó:

1 Siêu mặt phẳng trong dạng chuẩn phải thỏa mãn các ràng buộc:

Khoảng cách từ đến siêu mặt phẳng (w, b) được tính theo công thức:

Biên độ của siêu mặt phẳng (w, b) được tính theo công thức:

Vấn đề tối ưu trong SVM là việc tìm giá trị của w và b sao cho biên độ (margin) đạt giá trị lớn nhất và thỏa mãn ràng buộc:

Sau khi tìm ra đường phân cách siêu mặt phẳng , class cho dữ liệu mới sẽ được định nghĩa như sau:

Trong đó sgn là hàm định nghĩa kí hiệu, có giá trị = 1 nếu đối số không âm và -1 nếu đối số âm.

A computer screen shot of a program code

Description automatically generated

Xây dựng và huấn luyện mô hình SVM:

* **SVC:** Tạo một đối tượng SVM với kernel tuyến tính (kernel='linear'), giá trị C là 1 (điều chỉnh mức độ phạt cho các lỗi phân loại), và random\_state=42 để đảm bảo tính tái lập của kết quả.
* **fit:** Huấn luyện mô hình SVM với dữ liệu huấn luyện (X\_train\_vec, y\_train).

Dự đoán:

* **predict:** Sử dụng mô hình đã huấn luyện để dự đoán nhãn cho dữ liệu kiểm tra (X\_test\_vec).

Đánh giá mô hình:

* **accuracy\_score:** Tính độ chính xác của mô hình bằng cách so sánh nhãn dự đoán (y\_pred) với nhãn thực tế (y\_test).
* **classification\_report:** Tạo một báo cáo phân loại chi tiết bao gồm các chỉ số như precision, recall, F1-score cho từng lớp nhãn.

## 4.2 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) là một loại mô hình học sâu (deep learning) đặc biệt được thiết kế để xử lý dữ liệu dạng lưới, như hình ảnh. CNN được sử dụng rộng rãi trong các tác vụ liên quan đến xử lý hình ảnh và video, như nhận dạng đối tượng, CNN không chỉ được sử dụng trong xử lý hình ảnh mà còn có ứng dụng rộng rãi trong phân tích văn bản, bao gồm cả phân tích cảm xúc trên Twitter (Twitter Sentiment Analysis).

### 4.2.1 Kiến trúc CNN

* Lớp Embedding: Chuyển đổi các từ trong tweet thành các vector số học có kích thước cố định. Lớp này học các biểu diễn từ (word embeddings) trong quá trình huấn luyện.
* Lớp Convolution: Áp dụng các bộ lọc (filters) khác nhau để trích xuất các đặc trưng quan trọng từ các đoạn văn bản. Các bộ lọc này di chuyển qua văn bản để nhận diện các mẫu từ.
* Lớp Activation: Sử dụng các hàm kích hoạt phi tuyến như ReLU để thêm tính phi tuyến vào mô hình.
* Lớp Pooling: là Max Pooling, giúp giảm kích thước không gian của các đặc trưng và giữ lại các thông tin quan trọng nhất.
* Lớp Fully Connected: Kết nối tất cả các neuron từ lớp trước với mỗi neuron trong lớp này, thường là một hoặc nhiều lớp fully connected để tổng hợp các đặc trưng.
* Lớp Output: Lớp cuối cùng sử dụng hàm softmax để dự đoán xác suất của các lớp cảm xúc (ví dụ: positive, negative, neutral).

### 4.2.2 Huấn luyện và đánh giá

**Huấn luyện:**

* Loss Function: Sử dụng hàm mất mát (loss function) như categorical cross-entropy cho bài toán phân loại.
* Optimizer: Sử dụng các thuật toán tối ưu hóa như Adam hoặc SGD để cập nhật trọng số của mô hình.

**Đánh giá:**

* Validation: Đánh giá mô hình trên tập dữ liệu kiểm tra để theo dõi hiệu suất và điều chỉnh mô hình.
* Metrics: Sử dụng các chỉ số như accuracy, precision, recall, F1-score để đánh giá hiệu quả của mô hình.

A screenshot of a computer code

Description automatically generated

Xây dựng:

* **Embedding layer:** Chuyển đổi các chỉ số từ từ điển (với kích thước input\_dim=5000) thành các vector nhúng với kích thước embedding\_dim=100. Chiều dài đầu vào (input\_length) được xác định bởi max\_length.
* **Conv1D layer:** Áp dụng một lớp tích chập 1 chiều với 128 bộ lọc, kích thước kernel là 5, và hàm kích hoạt ReLU để phát hiện các đặc trưng cục bộ trong chuỗi đầu vào.
* **GlobalMaxPooling1D:** Lấy giá trị lớn nhất từ mỗi bộ lọc, giúp giảm chiều của dữ liệu và duy trì các đặc trưng quan trọng nhất.
* **Dense layer (hidden layer):** Một lớp fully-connected với 10 đơn vị và hàm kích hoạt ReLU.
* **Dropout:** Áp dụng dropout với tỷ lệ 0.5 để giảm overfitting.
* **Dense layer (output layer):** Lớp đầu ra với một đơn vị và hàm kích hoạt sigmoid để dự đoán xác suất của lớp nhãn (0 hoặc 1).

Biên dịch:

* **Loss:** Sử dụng hàm mất mát binary\_crossentropy cho nhiệm vụ phân loại nhị phân.
* **Optimizer:** Sử dụng thuật toán tối ưu Adam.
* **Metrics:** Sử dụng độ chính xác (accuracy) như là một thước đo hiệu quả của mô hình.

Thiết lập Early Stopping:

* **EarlyStopping:** Theo dõi giá trị val\_loss (mất mát trên tập dữ liệu kiểm tra), và nếu giá trị này không cải thiện sau 3 epochs, quá trình huấn luyện sẽ dừng lại và mô hình sẽ khôi phục lại trạng thái có trọng số tốt nhất.

Huấn luyện và đánh giá:

* **fit:** Huấn luyện mô hình với dữ liệu huấn luyện (X\_train\_pad, y\_train), trong tối đa 10 epochs, với kích thước batch là 64, và sử dụng dữ liệu kiểm tra (X\_test\_pad, y\_test) để đánh giá trong quá trình huấn luyện. Hàm callback early\_stop được sử dụng để thực hiện early stopping.
* **evaluate:** Đánh giá mô hình trên tập dữ liệu kiểm tra (X\_test\_pad, y\_test) và in ra độ chính xác của mô hình sau khi huấn luyện.

## 4.3 Recurrent Neural Network (RNN) - Long Short Term Memory (LSTM)

Recurrent Neural Network (RNN) là một loại mạng nơ-ron nhân tạo có khả năng xử lý dữ liệu dạng chuỗi. Điều này làm cho RNN trở thành một lựa chọn lý tưởng cho các nhiệm vụ liên quan đến xử lý ngôn ngữ tự nhiên (NLP), bao gồm phân tích sentiment trên các tweet từ Twitter.

Khác với các mạng nơ-ron truyền thống (như feedforward neural networks), RNN có khả năng ghi nhớ thông tin từ các bước trước đó trong chuỗi nhờ vào cơ chế "tái sử dụng" trạng thái ẩn (hidden state). Điều này cho phép RNN duy trì thông tin về ngữ cảnh trong chuỗi dữ liệu, giúp hiểu rõ hơn về mối quan hệ giữa các từ trong câu.

### 4.3.1 RNN

**Các dạng bài toán:**

A diagram of a number of rectangular objects

Description automatically generated with medium confidence

**One to one**: mẫu bài toán cho Neural Network (NN) và Convolutional Neural Network (CNN), 1 input và 1 output, ví dụ với CNN input là ảnh và output là ảnh được segment.

**One to many**: bài toán có 1 input nhưng nhiều output, ví dụ: bài toán caption cho ảnh, input là 1 ảnh nhưng output là nhiều chữ mô tả cho ảnh đấy, dưới dạng một câu.

**Many to one**: bài toán có nhiều input nhưng chỉ có 1 output, ví dụ bài toán phân loại hành động trong video, input là nhiều ảnh (frame) tách ra từ video, ouptut là hành động trong video

**Many to many**: bài toán có nhiều input và nhiều output, ví dụ bài toán dịch từ tiếng anh sang tiếng việt, input là 1 câu gồm nhiều chữ: “I love Vietnam” và output cũng là 1 câu gồm nhiều chữ “Tôi yêu Việt Nam”.

* **RNN có thể:** Dịch máy (Machine Translation), Mô hình hóa ngôn ngữ và sinh văn bản, Nhận dạng giọng nói, Mô tả hình ảnh: RNN kết hợp cùng CNN để sinh ra mô tả cho hình ảnh chưa được gán nhãn

**Kiến trúc RNN:**

* Lớp Embedding: Tương tự như CNN, chuyển đổi các từ trong tweet thành các vector số học.
* Lớp Recurrent:
* RNN Layer: Lớp này bao gồm các đơn vị (neurons) có khả năng duy trì trạng thái bên trong, cho phép mô hình nhớ thông tin từ các bước trước đó trong chuỗi. Có các loại RNN phổ biến như:
* Simple RNN: Dạng cơ bản của RNN, nhưng thường gặp vấn đề với vanishing gradient khi chuỗi dữ liệu dài.
* Long Short-Term Memory (LSTM): Một loại RNN đặc biệt, được thiết kế để giải quyết vấn đề vanishing gradient và có khả năng nhớ lâu hơn.
* Gated Recurrent Unit (GRU): Một biến thể của LSTM, đơn giản hơn và hiệu quả trong nhiều trường hợp.
* Lớp Fully Connected: Kết nối toàn bộ các neuron từ lớp RNN đến các lớp tiếp theo để tổng hợp các đặc trưng.
* Lớp Output: Lớp cuối cùng sử dụng hàm softmax để dự đoán xác suất của các lớp cảm xúc

**Ý tưởng căn bản:**

* Đối với mạng neural thông thường, chúng ta cho tất cả dữ liệu vào cùng một lúc. Nhưng đôi khi, dữ liệu của chúng ta mang ý nghĩa trình tự, tức nếu thay đổi trình tự dữ liệu, kết quả sẽ khác.
* Dễ thấy rõ nhất ở dữ liệu văn bản. Ví dụ, “Con ăn cơm chưa” và “Con chưa ăn cơm”, nếu tách mỗi câu theo từ, ta được bộ vocab [ ‘con’, ‘ăn’, ‘cơm’, ‘chưa’], one hot encoding và cho tất cả vào mạng neural , có thể thấy ngay, không có sự phân biệt nào giữa 2 câu trên. Việc đảo thứ tự duyệt các từ làm sai lệch ý nghĩ của câu.
* Nói cách khác, chúng ta cần một mạng neural có thể xử lí tuần tự:

A diagram of a network

Description automatically generated

Các x ở đây đại diện cho dữ liệu đầu vào lần lượt (được chia theo time step).

 đại diện cho time step thứ t, và ​ là output của một step. Ví dụ,  sẽ là vector đại diện của từ thứ 2 trong câu văn bản.

Hình ảnh dưới đây cho thấy rõ hơn điều gì thực sự xảy ra trong một step:

A diagram of a circuit

Description automatically generated

* **Hidden state** . Đây chính là bộ nhớ của mạng. ​ là tổng hợp thông tin của hidden state trước cộng với input tại time step *t*  Activation function ở đây chủ yếu là tanh hoặc ReLu.
* **Output của từng time step *​*** : Tại 1 block của mạng RNN có 2 đầu ra. Trong đó, ​ là tổng hợp thông tin của các state trước để tiếp tục truyền đi trong chuỗi mạng, và ta có thêm *​* là output của từng time step một. Ở đây *​* thường là hàm softmax.

)

**Tính toán lan truyền ngược (BPTT – Backpropagation Through Time):**

* Như vậy, trong quá trình training, có 3 tham số chúng ta cần tìm là ,,​. Chúng ta cần tính , ,    ​. (với *L* là loss function)

**A math equations and formulas

Description automatically generated**

* Trong mạng NN truyền thống, ta không chia sẻ tham số giữa các tầng mạng. Tuy vậy, với RNN, ta có thể thấy, để tính đạo hàm của loss theo ​, ta phụ thuộc vào ​, mà ​ lại phụ thuộc vào ​ và  ​. Nói nôm na, ta phải cộng tất cả đầu ra ở các bước trước để tính đạo hàm. Điều này gây ra một hạn chế lớn cho RNN.

### 4.3.2 LSTM

**RNN:**

A diagram of a diagram

Description automatically generated

**LSTM:**

A diagram of a diagram

Description automatically generated

Về cơ bản, ý tưởng không khác nhau là mấy. Chúng ta chỉ thêm một số tính toán ở đây. Tất cả được tóm tắt trong hình sau:

A black and white image of a square with a square with a square with a square with a square with a square with a square with a square with a square with a square with a square with

Description automatically generated

Đầu tiên, chúng ta có 𝑖,𝑓,𝑔*i*,*f*,*g* có công thức gần giống hệt nhau và chỉ khác mỗi ma trận tham số. Chính ma trận này sẽ quyết định chức năng khác nhau của từng cổng. 𝜎*σ* là ký hiệu của hàm sigmoid.

* **Input gate** 𝑖 - cổng vào.

Cổng vào giúp quyết định bao nhiêu lượng thông tin đầu vào sẽ ảnh hưởng đến trạng thái mới. Quyết định bằng cách nào, thông qua đặc điểm của hàm sigmoid (đầu ra nằm trong khoảng [0,1][0,1]), như vậy khi một vector thông tin đi qua đây, nếu nhân với 0, vector sẽ bị triệt tiêu hoàn toàn. Nếu nhân với 1, hầu hết thông tin sẽ được giữ lại.

* Tương tự như vậy, 𝑓là **forget gate** - cổng quên.

Cổng quyết định sẽ bỏ đi bao nhiêu lượng thông tin đến từ trạng thái trước đó.

* Cuối cùng, cổng 𝑜 là **output gate** - cổng ra.

Cổng điều chỉnh lượng thông tin có thể ra ngoài ​ và lượng thông tin truyền tới trạng thái tiếp theo.

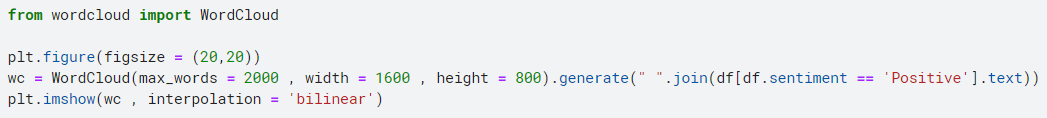
* Tiếp theo, g thực chất cũng chỉ là một trạng thái ẩn được tính dựa trên đầu vào hiện tại  và trạng thái trước ​. Tính hoàn toàn tương tự như input gate, chỉ thay vì dùng sigmoid, ta dùng tanh. Kết hợp hai điều này lại để cập nhật trạng thái mới.
* Cuối cùng, ta có ​ là bộ nhớ trong của LSTM. Nhìn vào công thức, có thể thấy nó là tổng hợp của bộ nhớ trước ​ đã được lọc qua *cổng quên f*, cộng với trạng thái ẩn g đã được lọc bởi *cổng vào i*. Cell state sẽ mang thông tin nào quan trọng truyền đi xa hơn và sẽ được dùng khi cần. Đây chính là **long term memory**.
* Sau khi có ​, ta sẽ đưa nó qua cổng ra để lọc thông tin một lần nữa, thu được trạng thái mới .

Nếu nhìn kỹ một chút, ta có thể thấy RNN truyền thống là dạng đặc biệt của LSTM. Nếu thay giá trị đầu ra của input gate là 1 và đầu ra forget gate là 0 (không nhớ trạng thái trước), ta được RNN thuần.

# CHƯƠNG 5: TRỰC QUAN HÓA

## 5.1 Trực quan hóa dữ liệu Keyword

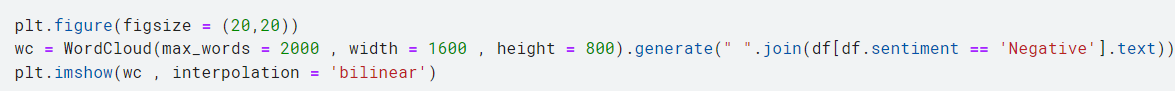
Trực quan hóa các keyword liên quan tới nhận xét tích cực:



A close-up of words

Description automatically generated

Trực quan hóa các keyword liên quan tới nhận xét tiêu cực:



A close up of words

Description automatically generated

## 5.2 Trực quan hóa bằng đồ thị

Đồ thị biểu thị *Model accuracy* và *Model loss* trong việc phân tích cảm xúc:

A screen shot of a computer code

Description automatically generated

A graph of a model

Description automatically generated with medium confidence

**Model accuracy** là tỷ lệ các dự đoán chính xác của mô hình trên tổng số dự đoán. Nó được tính toán bằng cách lấy số lượng dự đoán chính xác chia cho tổng số dự đoán.

**Model loss** là một giá trị số đo lường mức độ sai lệch giữa dự đoán của mô hình và giá trị thực tế. Nó được tính toán bằng cách sử dụng một hàm mất mát cụ thể, chẳng hạn như softmax cross-entropy hoặc mean squared error.

Đồ thị Confusion matrix (Ma trận nhầm lẫn):

A diagram of a negative and negative matrix

Description automatically generated with medium confidence

Dựa vào ma trận trên, ta có thể tính các chỉ số:

A screenshot of a computer

Description automatically generated

# CHƯƠNG 6: HƯỚNG NGHIÊN CỨU

## 6.1 Cải thiện mô hình

Thử nghiệm với các kiến trúc và siêu tham số khác nhau để cải thiện độ chính xác của mô hình:

* Tăng kích thước từ vựng và chiều dài chuỗi:
* Tăng input\_dim để mô hình có thể học được nhiều từ hơn.
* Tăng input\_length để mô hình có thể xử lý các chuỗi dài hơn.
* Sử dụng lớp Convolutional trước LSTM: Thêm một lớp Convolutional có thể giúp mô hình học được các đặc trưng không gian trước khi chuyển qua lớp LSTM.
* Sử dụng Bidirectional LSTM: Sử dụng LSTM hai chiều có thể giúp mô hình hiểu được ngữ cảnh từ cả hai phía của chuỗi.
* Điều chỉnh các tham số của LSTM: Tăng hoặc giảm số lượng đơn vị ẩn trong lớp LSTM.
* Thử nghiệm với các giá trị khác nhau của dropout và recurrent\_dropout.
* Thêm các lớp Dense: Thêm một hoặc nhiều lớp Dense có thể giúp mô hình học được các đặc trưng phức tạp hơn.
* Sử dụng callback để điều chỉnh học suất: Sử dụng LearningRateScheduler hoặc ReduceLROnPlateau để điều chỉnh học suất trong quá trình huấn luyện.

## 6.2 Mở rộng ứng dụng

1. Giám sát cảm xúc khách hàng theo thời gian thực

Mô tả: Bằng cách tích hợp phân tích cảm xúc với các công cụ giám sát thời gian thực, doanh nghiệp có thể theo dõi cảm xúc của khách hàng thông qua các tweet và phản hồi ngay lập tức khi có vấn đề phát sinh.

Ứng dụng:

* Phần mềm giám sát mạng xã hội: Tích hợp mô hình phân tích cảm xúc vào phần mềm giám sát mạng xã hội để theo dõi các tweet liên quan đến thương hiệu hoặc sản phẩm của công ty theo thời gian thực.
* Phản hồi tự động: Thiết lập hệ thống phản hồi tự động gửi tin nhắn hỗ trợ hoặc giải quyết vấn đề khi phát hiện các tweet có cảm xúc tiêu cực..

# CHƯƠNG 6: TÀI LIỆU THAM KHẢO

[1]<https://ir.vnulib.edu.vn/>

[2][https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/#cnn-convolutional-neural-network-la-gi](https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/%23cnn-convolutional-neural-network-la-gi)

[3]<https://nttuan8.com/bai-13-recurrent-neural-network/#Recurrent_Neural_Network_la_gi>

[4] <https://arxiv.org/pdf/1601.06971>

[5] <https://viblo.asia/p/recurrent-neural-network-tu-rnn-den-lstm-gGJ597z1ZX2>

**—HẾT—**