|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THủy LỢI



NGUYỄN ĐỨC ANH

**Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử**

ĐỒ ÁN TỐT NGHIỆP

HÀ NỘI, NĂM 2026

|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THủy LỢI

NGUYỄN ĐỨC ANH

**Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử**

Ngành: Công nghệ thông tin Mã số:

NGƯỜI HƯỚNG DẪN: TS.Nguyễn Mạnh Hiển

HÀ NỘI, NĂM 2026

|  |  |
| --- | --- |
| https://lh7-us.googleusercontent.com/nR-016uA6aan1gCN6AHyUEwmR5lKdE8YE8jNmQ3rkSQC3kmnjzWJ4noql5yXHXE-m9HbGCEKL3JWVhXDNS5kzQDDQ9RB18SVLIy7WCDwSLYztON_AlpAbHunhPK3YWDRJMnjK0WbnVvIP6U0zYKC | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA** **VIỆT NAM**  **Độc lập - Tự do - Hạnh phúc**  ----------★----------  **NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP** |

**Họ tên sinh viên:** NGUYỄN ĐỨC ANH **Hệ đào tạo:** Đại học chính quy

**Lớp:**  63CNTT4 **Ngành:**  Công nghệ thông tin

**Khoa:**  Công nghệ thông tin

1. **TÊN ĐỀ TÀI: PHÂN TÍCH CẢM XÚC KHÁCH HÀNG TRONG CÁC BÌNH LUẬN SẢN PHẨM TRÊN CÁC SÀN THƯƠNG MẠI ĐIỆN TỬ**

**2- NỘI DUNG CÁC PHẦN THUYẾT MINH VÀ TÍNH TOÁN:**                 Tỷ lệ %

|  |  |
| --- | --- |
| **Nội dung các phần** | **Tỷ lệ %** |
| Chương 1: Tổng quan về hệ thống quản lý văn bản và quy trình xét duyệt văn bản trong các tổ chức sinh viên | 20% |
| Chương 2: Phân tích và thiết kế hệ thống | 40% |
| Chương 3: Xây dựng và cài đặt hệ thống | 40% |

**3. GIÁO VIÊN HƯỚNG DẪN TỪNG PHẦN**

|  |  |
| --- | --- |
| **Phần** | **Họ và tên giáo viên hướng dẫn** |
| Chương 1: Tổng quan về hệ thống quản lý văn bản và quy trình xét duyệt văn bản trong các tổ chức sinh viên | TS. Nguyễn Mạnh Hiển |
| Chương 2: Phân tích và thiết kế hệ thống |
| Chương 3: Xây dựng và cài đặt hệ thống |

**4. NGÀY GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP**

Ngày ............  tháng .........  năm 2025

|  |  |
| --- | --- |
| **Trưởng Bộ môn**  *(Ký và ghi rõ Họ tên)* | **Giáo viên hướng dẫn chính**  *(Ký và ghi rõ Họ tên)* |

Nhiệm vụ Đồ án tốt nghiệp đã được Hội đồng thi tốt nghiệp của Khoa thông qua.

                                     Ngày. . . . .tháng. . . . .năm 2025

**Chủ tịch Hội đồng**

*(Ký và ghi rõ Họ tên)*

Sinh viên đã hoàn thành và nộp bản Đồ án tốt nghiệp cho Hội đồng thi ngày..... tháng... năm 2025

**Sinh viên làm Đồ án tốt nghiệp**

*(Ký và ghi rõ Họ tên)*

|  |  |
| --- | --- |
| logo | TRƯỜNG ĐẠI HỌC THUỶ LỢI  **KHOA CÔNG NGHỆ THÔNG TIN**  BẢN TÓM TẮT ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP |

**Tên đề tài: Phân tích cảm xúc của khách hàng trong các bình luận sản phẩm trên các sàn thương mại điện tử**

*Sinh viên thực hiện:* Nguyễn Đức Anh

*Lớp:* 63CNTT4

*Mã sinh viên:* 2151062703

*Số điện thoại:* 0978898633

*Email:* [2151062703@e.tlu.edu.vn](mailto:2151062703@e.tlu.edu.vn)

*Giảng viên hướng dẫn:* TS. Nguyễn Mạnh Hiển

TÓM TẮT ĐỀ TÀI

Đề tài nghiên cứu phương pháp và xây dựng hệ thống phân tích cảm xúc (Sentiment Analysis) cho các bình luận sản phẩm trên các sàn thương mại điện tử (ví dụ: Shoppee, Lazada, Tiki). Mục tiêu của đề tài là thu thập dữ liệu đánh giá/bình luận của người dùng, thực hiện tiền xử lý văn bản tiếng việt, xây dựng và so sánh các mô hình phân loại cảm xúc(nhị phân, tam phân hoặc đa lớp), đánh giá hiệu năng mô hình bằng các chỉ số phù hợp, đồng thời triển khai một prototype giúp doanh nghiệp theo dõi mức độ hài long khách hàng theo sản phâm và theo thời gian. Đề tài kết hợp các kỹ thuật xử lý ngôn ngữ tự nhiên(NLP), học máy (Machine Learning) và thực tiễn triển khai phần mềm.

Nhận diện cảm xúc trong văn bản là một trong những bài toán quan trọng trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP), đặc biệt trong bối cảnh dữ liệu văn bản ngày càng gia tăng từ các nền tảng mạng xã hội và phương tiện truyền thông số. Đề tài này tập trung nghiên cứu và phát triển mô hình học máy để nhận diện cảm xúc trong văn bản tiếng Việt, dễ mở rộng và hiệu quả cao đối với dữ liệu văn bản ngắn như bình luận thương mại điện tử.

Cụ thể bình luận tiếng Việt sau khi được làm sạch và chuẩn hóa sẽ được biểu diễn dưới dạng đặc trưng số bằng các phương pháp biểu diễn văn bản phổ biến như Bag-of-Words(BoW) và TF-IDF (Term Frequency-Inverse Document Frequency), có thể kết hợp n-gram để khai thác các cụm từ quan trọng trong cảm xúc(ví dụ: “không tốt”, “rất hài lòng”,” “giao hàng chậm”). Trên cơ sở đặc trưng này, đề tài sử dụng mô hình Support Vector Machine(SVM) để phân loại văn bản nhờ khả năng xử lý tốt dữ liệu thưa và có số chiều lớn (đặc trưng TF-IDF/BoW), đồng thời có thể điều chỉnh bằng tham số C và lựa chọn kernel phù hợp tối ưu hiệu năng.

Bằng cách thử nghiệm và so sánh các tổ hợp phương pháp (BoW + SVM, TF-IDF + SVM, n-gram TF-IDF + SVM), đề tài hướng đến xây dựng mô hình phân loại cảm xúc hiệu quả và có tính ứng dụng cao trong các bài toán như phân tích phản hồi khách hàng, đánh giá sản phẩm, và hỗ trợ doanh nghiệp theo dói xu hương hài lòng của người dùng trên sàn thương mại điện tử.

CÁC MỤC TIÊU CHÍNH

1. Nghiên cứu tổng quan về phân tích cảm xúc trong văn bản.
2. Thu thập và tiền xử lý tập dữ liệu cảm xúc tiếng Việt từ các nguồn đáng tin cậy.
3. Xây dựng và thử nghiệm các mô hình học máy.
4. Đánh giá và so sánh hiệu suất mô hình.
5. Đề xuất hướng cải tiến và ứng dụng thực tế.

KẾT QUẢ DỰ KIẾN

* **Xây dựng tập dữ liệu cảm xúc tiếng Việt**: Thu thập và tiền xử lý một bộ dữ liệu văn bản tiếng Việt có gán nhãn cảm xúc, phục vụ cho quá trình huấn luyện và đánh giá mô hình.
* **Phát triển và triển khai các mô hình học máy**: Áp dụng các phương pháp như Naïve Bayes, Recurrent Neural Network (RNN) với Long Short-Term Memory (LSTM), kết hợp với các kỹ thuật vector hóa từ như Word2Vec (CBOW) và TF-IDF để phân loại cảm xúc.
* **Đánh giá hiệu suất mô hình**: So sánh độ chính xác và hiệu quả của các mô hình đã triển khai, từ đó đề xuất mô hình tối ưu nhất cho bài toán nhận diện cảm xúc trong văn bản tiếng Việt.

LỜI CAM ĐOAN

Em xin cam đoan đồ án tốt nghiệp với đề tài “Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình học máy” là công trình nghiên cứu do chính em thực hiện dưới sự hướng dẫn của TS. Nguyễn Mạnh Hiển. Các nội dung trình bày trong đồ án là sản phẩm nghiên cứu, tổng hợp và thực nghiệm của cá nhân em; mọi số liệu, kết quả đều trung thực và chưa từng được công bố trong bất kỳ công trình nào khác.

Các tài liệu tham khảo được sử dụng đều có nguồn gốc rõ ràng và được trích dẫn theo đúng quy định. Em hoàn toàn chịu trách nhiệm về tính trung thực và bản quyền của đồ án này.

|  |  |
| --- | --- |
|  | **Tác giả ĐATN**  *Chữ ký*  **Trương Huy Phú** |

LỜI CẢM ƠN

Bốn năm là khoảng thời gian không quá dài nhưng cũng chả ngắn đối với thanh xuân của một đời người. Bốn năm ấy là bốn năm với biết bao bài học, bài học từ thành công, bài học từ thất bại, chung quy lại đều là những thứ em rất trân quý. Bốn năm bản thân em được gắn bó và đồng hành cùng Khoa Công nghệ thông tin, trường Đại học Thủy Lợi, nhận được sự quan tâm và hướng dẫn tận tình từ thầy cô giảng dạy tại ngôi trường thân yêu, được kết nối đến các bạn, các anh chị, đấy không chỉ là tri thức mà còn là kỷ niệm, và trải nghiệm. Không những được học tập, mà em còn được cọ xát, học hỏi từ các cuộc thi như Olympic Toán học, ca múa hát, … Nhờ những điều ấy mà em được rèn giũa và trưởng thành hơn. Em xin gửi lời cảm ơn chân thành đến các thầy cô giáo đã giảng dạy và tạo điều kiện thuận lợi trong quá trình em học tập tại trường.

Em xin bày tỏ sự trân trọng, lòng biết ơn và lời cảm ơn sâu sắc nhất đến thầy TS. Nguyễn Mạnh Hiển, giảng viên bộ môn Mạng và An toàn thông tin, Khoa công nghệ thông tin, trường Đại học Thủy Lợi. Sự tận tình chỉ bảo, hướng dẫn và nhiệt huyết của thầy đã tạo nguồn động lực to lớn cho em, thúc đẩy em không ngừng cố gắng để hoàn thành Đồ án tốt nghiệp một cách chỉnh chu và hoàn thiện nhất.

Xin cảm ơn những người đã luôn đồng hành cùng với em trên con đường vừa qua. Cảm ơn gia đình, bạn bè, các anh chị đã luôn giúp đỡ, động viên và ủng hộ em trong quá trình học tập.

Đồ án tốt nghiệp được em cố gắng thực hiện và hoàn thành với tất cả sự nỗ lực của bản thân nhưng không thể tránh khỏi được một số thiếu sót, em rất mong sẽ nhận được sự góp ý từ thầy cô, bạn bè và những người quan tâm tới đề tài này để đồ án được hoàn thiện hơn nữa.

Em xin chân thành cảm ơn!

MỤC LỤC

[DANH MỤC HÌNH ẢNH iv](#_Toc201869521)

[DANH MỤC BẢNG vii](#_Toc201869522)

[DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ viii](#_Toc201869523)

[CHƯƠNG 1 : TỔNG QUAN VỀ TÀI LIỆU 5](#_Toc201869524)

[1.1 Ngôn ngữ tự nhiên 5](#_Toc201869525)

[1.2 Ngôn ngữ tiếng Việt 6](#_Toc201869526)

[1.3 Xử lý ngôn ngữ tự nhiên 8](#_Toc201869527)

[CHƯƠNG 2 CƠ SỞ LÝ THUYẾT 10](#_Toc201869528)

[2.1 Các mô hình mạng nơ-ron sử dụn trong học sâu 10](#_Toc201869529)

[2.1.1 Khái niệm tổng quát 10](#_Toc201869530)

[2.1.2 Một số kiến trúc mạng nơ-ron phổ biến 11](#_Toc201869531)

[2.2 Biểu diễn văn bản bằng Word2Vec 12](#_Toc201869532)

[2.2.1 Khái quát chung 12](#_Toc201869533)

[2.2.2 Nguyên lý hoạt động 13](#_Toc201869534)

[2.2.3 Các kiến trúc mô hình 13](#_Toc201869535)

[2.2.4 Thuật toán tối ưu 15](#_Toc201869536)

[2.3 Các mô hình nhận diện cảm xúc trong văn bản 16](#_Toc201869537)

[2.3.1 Tiếp cận phân tích cảm xúc từ xử lý ngôn ngữ tự nhiên 16](#_Toc201869538)

[2.3.2 Tiếp cận phân tích cảm xúc bằng phương pháp học máy 16](#_Toc201869539)

[2.3.3 Kiến trúc tổng quát của mô hình 17](#_Toc201869540)

[CHƯƠNG 3 NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT 19](#_Toc201869541)

[3.1 Tiền xử lý dữ liệu văn bản 19](#_Toc201869542)

[3.1.1 Tách từ 19](#_Toc201869543)

[3.1.2 Chuẩn hóa từ ngữ và chính tả 21](#_Toc201869544)

[3.2 Vector hóa văn bản 22](#_Toc201869545)

[3.2.1 Các phương pháp biểu diễn văn bản cổ điển (Classic Word Embedding) 22](#_Toc201869546)

[3.2.2 Các phương pháp biểu diễn văn bản hiện đại 25](#_Toc201869547)

[3.3 Mô hình nhận diện cảm xúc sử dụng học sâu và học máy 26](#_Toc201869548)

[CHƯƠNG 4 THỰC NGHIỆM 28](#_Toc201869549)

[4.1 Xây dựng dữ liệu 28](#_Toc201869550)

[4.1.1 Bối cảnh và vai trò của dữ liệu huấn luyện 28](#_Toc201869551)

[4.1.2 Bộ dữ liệu nhận xét được gán nhãn cảm xúc tích cực, tiêu cực 28](#_Toc201869552)

[4.2 Tiền xử lý dữ liệu 31](#_Toc201869553)

[4.3 Huấn luyện mô hình 32](#_Toc201869554)

[4.3.1 Mô hình Naive Bayes và SVM với TF-IDF 33](#_Toc201869555)

[3.3.2 Mô hình học sâu LSTM với Word2Vec 34](#_Toc201869556)

[KẾT LUẬN 37](#_Toc201869557)

[TÀI LIỆU THAM KHẢO 38](#_Toc201869558)

[PHỤ LỤC 39](#_Toc201869559)

DANH MỤC HÌNH ẢNH

[Hình 2.1 Cách biểu diễn các từ trên Word2Vec 12](#_Toc202088899)

[Hình 2.2 Mô hình CBOW 14](#_Toc202088900)

[Hình 2.3 Mô hình Skip-Gram 15](#_Toc202088901)

[Hình 3.1 Mô hình biểu diễn Bag of Words 23](#_Toc202088902)

[Hình 3.2 Ma trận biểu diễn theo phương pháp Distributional Embedding 25](#_Toc202088903)

[Hình 3.3 Mô hình CBOW và Skip-Gram 26](#_Toc202088904)

[Hình 3.4 Mô hình SAV 27](#_Toc202088905)

[Hình 4.1 Cấu trúc thư mục data\_train, và data\_test 29](#_Toc202088906)

[Hình 4.2 Mô tả dữ liệu của một file 30](#_Toc202088907)

[Hình 4.3 Sơ đồ huấn luyện 33](#_Toc202088908)

[Hình 4.4 Sơ đồ kiểm tra 35](#_Toc202088909)

[Hình 4.5 Báo cáo trên tập dữ liệu kiểm tra với PP LSTM 39](#_Toc202088910)

[Hình 4.6 Báo cáo trên tập dữ liệu kiểm tra với PP SVM 41](#_Toc202088911)

[Hình 4.7 Báo cáo trên tập dữ liệu kiểm tra với PP Naïve Bayes 41](#_Toc202088912)

DANH MỤC BẢNG

[Bảng 4.1 Cấu trúc bộ dữ liệu data\_train 30](#_Toc202088968)

[Bảng 4.2 Cấu trúc bộ dữ liệu data\_test 31](#_Toc202088969)

[Bảng 4.3 Danh sách các mô hình vector hóa và thuật toán phân loại được sử dụng trong thực nghiệm 38](#_Toc202088970)

[Bảng 4.4 Các lớp chính trong mô hình LTSMs 38](#_Toc202088971)

[Bảng 4.5 Kết quả mỗi epoch huấn luyện 39](#_Toc202088972)

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

**ĐATN** Đồ án tốt nghiệp

**IEEE** Institute of Electrical and Electronics Engineers

**KLTN** Khóa luận tốt nghiệp

**LVTN** Luận văn tốt nghiệp

**SAV** Sentiment Analysis Vietnamese

**RNN** Recurrent Neural Network

**LSTMs** Long Short Term Memory Neural Network

**Word2Vector** vector representations of words

**CBOW** Continuous Bag-of-Words

**NLP** Natural Language Processing

MỞ ĐẦU

Trong bối cảnh các lĩnh vực công nghệ ngày càng phát triển mạnh mẽ, việc tự động nhận diện cảm xúc trong văn bản tiếng Việt đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực như: quản trị doanh nghiệp, xây dựng và phát triển thương hiệu, chăm sóc khách hàng, khảo sát ý kiến và phân tích đánh giá phản hồi từ người dùng.

Ý kiến và cảm nhận của khách hàng hiện nay đóng vai trò vô cùng quan trọng trong việc định hướng sản phẩm cũng như chiến lược kinh doanh. Do đó, các doanh nghiệp ngày càng có nhu cầu cao trong việc triển khai các hệ thống có khả năng phân tích phản hồi khách hàng một cách tự động và hiệu quả. Thông qua đó, doanh nghiệp có thể nắm bắt được thị hiếu, xu hướng và kỳ vọng của người tiêu dùng, từ đó kịp thời điều chỉnh sản phẩm, nâng cao năng lực cạnh tranh, cũng như thích ứng tốt hơn với những thay đổi liên tục của thị trường.

Từ góc độ nghiên cứu, việc xây dựng một hệ thống có khả năng phân tích cảm xúc trong văn bản tiếng Việt là một trong những hướng đi quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Điều này không chỉ góp phần giải quyết các bài toán thực tiễn mà còn thúc đẩy tiến bộ kỹ thuật trong việc xử lý và hiểu ngữ nghĩa tiếng Việt.

Trong đề tài này, chúng tôi tập trung vào việc xây dựng mô hình phân loại cảm xúc người dùng dựa trên văn bản đánh giá và phản hồi. Cụ thể, các cảm xúc sẽ được chia thành hai nhóm chính và bài toán được triển khai dưới dạng một bài toán phân lớp. Mỗi phản hồi được biểu diễn dưới dạng vector đặc trưng để đưa vào mô hình học máy, nhằm huấn luyện hệ thống có khả năng nhận diện và phân loại cảm xúc của người dùng một cách chính xác.

**Tổng quan về vấn đề nghiên cứu**

Trong thời gian gần đây, bài toán phân tích và nhận diện cảm xúc đã trở nên phổ biến hơn khi được ứng dụng rộng rãi vào việc xử lý khối lượng lớn dữ liệu trên các nền tảng truyền thông xã hội như mạng xã hội, diễn đàn trực tuyến, blog cá nhân, wiki, hay các hệ thống cộng tác trực tuyến. Đây là một lĩnh vực quan trọng trong điện toán cảm xúc, với mục tiêu chính là phân loại văn bản (và mở rộng sang các dạng dữ liệu khác như âm thanh, hình ảnh) theo chiều hướng cảm xúc — thông thường là tích cực hoặc tiêu cực.

Bài toán này có mối liên hệ chặt chẽ với các lĩnh vực như truy xuất và tổng hợp thông tin, vì cần thực hiện nhiều bước: thu thập dữ liệu, xử lý và gán nhãn trước khi đưa vào mô hình phân tích. Dù phần lớn các nghiên cứu hiện nay tập trung vào ngôn ngữ tiếng Anh, số lượng công trình nghiên cứu mở rộng sang các ngôn ngữ khác, bao gồm cả tiếng Việt, đang ngày càng gia tăng.

Hệ thống nhận diện cảm xúc hiện có thể chia thành hai nhóm chính dựa trên cách tiếp cận: dựa trên tri thức (knowledge-based) và dựa trên thống kê (statistical-based). Trong nhiều nghiên cứu, bài toán này thường được đơn giản hóa thành bài toán phân lớp. Tuy nhiên, trên thực tế, đây là một nhiệm vụ phức tạp thuộc lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP), đòi hỏi phải kết hợp nhiều kỹ thuật hỗ trợ như: nhận diện thực thể có tên, trích xuất khái niệm, phát hiện sự châm biếm, phân tích các khía cạnh, và đặc biệt là phát hiện tính chủ quan trong ngữ cảnh. Việc xác định tính chủ quan là một bước không thể thiếu, bởi vì đa số mô hình hiện nay được thiết kế để phân biệt giữa những phát biểu mang tính cảm xúc và các phát biểu mang tính trung lập.

Đối với tiếng Việt, cộng đồng nghiên cứu hiện mới chỉ giải quyết hiệu quả bài toán ở mức đơn giản — cụ thể là phân loại văn bản thành hai nhóm cảm xúc cơ bản là tích cực và tiêu cực, với độ chính xác trung bình đạt khoảng 85% trở lên tùy theo mô hình và tập dữ liệu sử dụng.

Một số phương pháp tiêu biểu được áp dụng để xử lý bài toán phân tích cảm xúc bao gồm:

• Phương pháp thủ công (Keyword-based):

Đây là cách tiếp cận truyền thống, dựa vào việc xác định các từ mang sắc thái cảm xúc trong văn bản. Từng từ được gán điểm số tích cực hoặc tiêu cực và sau đó được tổng hợp lại để xác định cảm xúc tổng thể. Mặc dù dễ triển khai và tốc độ xử lý nhanh, phương pháp này lại không quan tâm đến ngữ cảnh hoặc trật tự từ, dẫn đến độ chính xác thấp nếu bộ từ điển không đủ mạnh. Ngoài ra, các từ phủ định hoặc mang nghĩa chuyển (ẩn dụ, mỉa mai) có thể khiến kết quả bị sai lệch.

• Phương pháp học sâu (Deep Learning):

Một trong những hướng tiếp cận hiện đại là sử dụng các mô hình học sâu như Recurrent Neural Network (RNN), trong đó mô hình Long Short-Term Memory (LSTM) được sử dụng phổ biến do khả năng ghi nhớ chuỗi dài và xử lý thông tin ngữ cảnh tốt. Kết hợp với đó là kỹ thuật biểu diễn từ bằng vector (Word Embedding) như Word2Vec, đặc biệt với kiến trúc CBOW (Continuous Bag-of-Words) giúp ánh xạ từ ngữ sang không gian vector liên tục nhằm phản ánh quan hệ ngữ nghĩa giữa các từ.

• Phương pháp kết hợp luật và thống kê (Rule-based + Corpus-based):

Phương pháp này kết hợp giữa mô hình học máy sâu như Recursive Neural Network (RNN) với tri thức ngôn ngữ chuyên gia thông qua cấu trúc cây cú pháp — điển hình là Sentiment Treebank. Trong mô hình cây này, mỗi nút biểu diễn một cụm từ hoặc câu và được gán nhãn cảm xúc theo các mức độ: rất tiêu cực, tiêu cực, trung lập, tích cực và rất tích cực. Mặc dù cách tiếp cận này cho kết quả khá tốt trong tiếng Anh, việc áp dụng sang tiếng Việt còn nhiều hạn chế do thiếu tập dữ liệu tương ứng và công cụ phân tích cú pháp chuyên biệt cho tiếng Việt.

**Mục đích nghiên cứu**

Mục tiêu chính của đề tài là nghiên cứu và xây dựng một mô hình có khả năng nhận diện cảm xúc trong văn bản tiếng Việt, cụ thể là từ các đánh giá, phản hồi của người dùng. Cảm xúc trong văn bản được phân loại theo hai tiêu chí định tính cơ bản:

* Xác định mức độ **tích cực** hoặc **tiêu cực** của nội dung.
* Phân biệt giữa văn bản mang tính **chủ quan** và **khách quan**.

Ngoài ra, đề tài cũng hướng đến việc tối ưu hiệu năng của mô hình cả về **độ chính xác** và **thời gian xử lý**, từ đó góp phần nâng cao hiệu quả trong các ứng dụng phân tích cảm xúc của người dùng, cũng như đóng góp vào lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt.

**Đối tượng và phạm vi nghiên cứu**

* **Đối tượng nghiên cứu**: Các văn bản tiếng Việt thể hiện cảm xúc, đặc biệt là các phản hồi, đánh giá của người dùng về sản phẩm, dịch vụ. Từ các đặc điểm của văn bản, mô hình sẽ học để xác định cảm xúc tương ứng.
* **Phạm vi nghiên cứu**: Dữ liệu sử dụng để huấn luyện và đánh giá mô hình được thu thập từ các bình luận của trang web [foody.vn](https://www.foody.vn/).

**Phương pháp nghiên cứu**

Đồ án kết hợp giữa nghiên cứu lý thuyết và xây dựng mô hình thực nghiệm:

* **Tìm hiểu và tổng hợp tài liệu** liên quan đến chủ đề như: các nghiên cứu trước đây, mô hình phân tích cảm xúc, các kỹ thuật học máy và xử lý ngôn ngữ tự nhiên.
* **Về mặt lý thuyết**: Nghiên cứu tổng quan về lĩnh vực phân tích cảm xúc trong văn bản tiếng Việt, các phương pháp phổ biến trong nhận diện cảm xúc và một số mô hình tiên tiến được ứng dụng trong các công trình khoa học.
* **Về thực nghiệm**: Tiến hành huấn luyện và đánh giá các mô hình học máy trên tập dữ liệu thực tế; sử dụng công cụ lập trình để xử lý, phân tích, và đánh giá độ chính xác dựa trên hai loại cảm xúc chính là **tích cực** và **tiêu cực**. Qua đó rút ra nhận xét và đánh giá hiệu quả của từng mô hình cũng như các tham số kỹ thuật đi kèm.

**Kết cấu của đề tài**

Nội dung đề tài bao gồm các chương:

**Mở đầu:** Giới thiệu đề tài

**Chương 1:** Tổng quan về nền tảng và công nghệ

**Chương 2:** Phân tích và thiết kế hệ thống.

**Chương 3:** Xây dựng và cài đặt hệ thống.

**Chương 4:** Kết luận và hướng phát triển

**Phụ lục:** Mã nguồn, giao diện chính

**Tài liệu tham khảo**

# TỔNG QUAN VỀ TÀI LIỆU

## Ngôn ngữ tự nhiên

Trong lĩnh vực ngôn ngữ học, “ngôn ngữ tự nhiên” được hiểu là các loại ngôn ngữ hình thành một cách tự nhiên trong quá trình tiến hóa và giao tiếp của loài người, không phải là kết quả của sự lập trình hay thiết kế có chủ đích. Đây là phương tiện giao tiếp cơ bản giữa con người với nhau và có thể tồn tại dưới nhiều hình thức khác nhau như lời nói, chữ viết, ngôn ngữ ký hiệu hay các dạng cảm nhận xúc giác.

Hiểu một cách đơn giản, ngôn ngữ tự nhiên (Natural Language) chính là các ngôn ngữ như tiếng Việt, tiếng Anh, tiếng Pháp... được con người sử dụng hàng ngày, khác biệt hoàn toàn so với các ngôn ngữ nhân tạo như ngôn ngữ lập trình (Python, Java, Pascal...) hay hệ thống mã hóa như mã Morse hoặc Braille.

Theo các thống kê hiện nay, có khoảng 5.600 ngôn ngữ được sử dụng trên toàn thế giới, tuy nhiên số lượng ngôn ngữ có hệ thống chữ viết chính thức thì chiếm tỷ lệ rất nhỏ.

**Đặc điểm của ngôn ngữ tự nhiên:**

* Ngôn ngữ tự nhiên là hiện tượng xã hội đặc biệt, phản ánh tư duy, văn hóa và lịch sử của cộng đồng người sử dụng.
* Đây là phương tiện truyền đạt thông tin phổ biến và quan trọng nhất trong xã hội loài người.
* Ngôn ngữ tự nhiên vận hành như một hệ thống tín hiệu có quy tắc, có thể phân tích về mặt ngữ âm, ngữ pháp và ngữ nghĩa.

**Phân loại ngôn ngữ tự nhiên:**

Ngôn ngữ tự nhiên có thể được phân chia theo nhiều tiêu chí:

* **Theo nguồn gốc lịch sử**: phân chia dựa trên quá trình phát triển và quan hệ giữa các ngôn ngữ.
* **Theo trật tự từ**: dựa vào cấu trúc câu như trật tự SVO (Chủ ngữ - Động từ - Tân ngữ).
* **Theo loại hình học ngôn ngữ**: phân chia dựa trên đặc điểm cấu trúc hình thái và chức năng. Mỗi loại hình sẽ có những thuộc tính phổ quát (chung), đặc trưng riêng, và thuộc tính phân loại để phân biệt giữa các ngôn ngữ.

## Ngôn ngữ tiếng Việt

Tiếng Việt thuộc loại hình ngôn ngữ đơn lập, nghĩa là các âm tiết được phát âm tách biệt, mỗi âm tiết thường tương ứng với một đơn vị chữ viết. Đặc trưng này ảnh hưởng đến hầu hết các khía cạnh của ngôn ngữ như âm vị, từ vựng và cú pháp.

**Về ngữ âm:**

Trong tiếng Việt, đơn vị cơ bản được gọi là “tiếng”, đồng thời cũng là đơn vị ngữ âm, mỗi tiếng tương ứng với một âm tiết. Hệ thống âm vị tiếng Việt phong phú và được tổ chức cân đối. Đặc biệt, nhiều từ ngữ trong tiếng Việt mang tính tượng thanh, tượng hình cao, giúp tăng khả năng biểu đạt cảm xúc và hình ảnh. Việc sử dụng ngôn ngữ đòi hỏi sự chú ý đến âm điệu và sự hài hòa giữa các tiếng trong câu.

**Về từ vựng:**

Trong tiếng Việt, mỗi tiếng đều mang ý nghĩa nhất định, và có thể kết hợp để tạo thành các đơn vị từ vựng mới thông qua hai phương thức chính: **ghép từ** và **láy âm**.

* **Từ ghép**: kết hợp hai hay nhiều tiếng theo quy tắc ngữ nghĩa (ví dụ: “đất nước”, “nhà lầu xe hơi”) hoặc có thể vay mượn và biến đổi theo thời đại (ví dụ: “thư điện tử” (e-mail), “phiên bản” , “truy cập ngẫu nhiên”).
* **Từ láy**: hình thành nhờ quy luật lặp lại hoặc biến đổi âm thanh để tạo hiệu ứng ngữ nghĩa (ví dụ: “long lanh”, “đom đóm”, “ầm ầm”, “lấm tấm” ...).

**Về ngữ pháp:**

Tiếng Việt là ngôn ngữ không biến hình, nghĩa là từ không thay đổi hình thức khi chuyển ngữ nghĩa hay vai trò trong câu. Do đó, cú pháp trong tiếng Việt phụ thuộc mạnh vào **trật tự từ** và **các hư từ** (như là, thì, đã, cũng, không...).

* **Trật tự từ**: đóng vai trò then chốt trong việc xác định nghĩa câu. Ví dụ: “Mùa xuân lại đến” khác hoàn toàn về ý nghĩa với “Lại đến mùa xuân”, dù đều gồm các từ giống nhau.
* **Hư từ**: giúp biểu đạt mối quan hệ logic, mức độ, sắc thái cảm xúc giữa các thành phần trong câu. Thay đổi vị trí hoặc loại hư từ có thể làm thay đổi sắc thái biểu cảm của cả câu. Ví dụ:
  + “Bạn ấy không uống nước ngọt.”
  + “Nước ngọt, bạn ấy không uống.”
  + “Nước ngọt, bạn ấy cũng không uống.”

## Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một lĩnh vực thuộc ngành khoa học máy tính, kết hợp giữa trí tuệ nhân tạo và ngôn ngữ học tính toán. Mục tiêu chính của NLP là giúp máy tính có khả năng hiểu, diễn giải và tương tác với ngôn ngữ con người một cách tự nhiên nhất, cho dù đầu vào là văn bản hay tiếng nói.

Trong bối cảnh dữ liệu ngày càng đa dạng, văn bản và âm thanh đang trở thành hai dạng dữ liệu phổ biến và được lưu trữ nhiều dưới dạng số. Tuy nhiên, đặc điểm của dữ liệu ngôn ngữ là thiếu cấu trúc hoặc chỉ có cấu trúc một phần, khiến chúng không thể xử lý bằng các phương pháp truyền thống như bảng biểu. Vì vậy, cần có những kỹ thuật đặc thù để chuyển các dạng dữ liệu này thành thông tin có thể xử lý tự động được.

Xử lý ngôn ngữ tự nhiên đóng vai trò như một cây cầu nối giữa con người và máy móc, giúp máy tính tham gia vào các công việc liên quan đến ngôn ngữ như dịch tự động, phân tích văn bản, hiểu và phản hồi lời nói, tóm tắt thông tin, truy vấn dữ liệu, ...

**Các bài toán tiêu biểu trong NLP:**

* **Nhận dạng giọng nói (Speech Recognition)**: Đây là công nghệ cho phép máy tính hiểu và chuyển lời nói thành văn bản. Nhiều trợ lý ảo như Siri (Apple), Google Assistant, Alexa (Amazon) hay Cortana (Microsoft) đều sử dụng kỹ thuật này để tương tác với người dùng.
* **Tổng hợp giọng nói (Text-to-Speech – TTS)**: Ngược lại với nhận dạng giọng nói, công nghệ này chuyển đổi văn bản thành âm thanh để thiết bị có thể "nói chuyện" với người dùng. Dù đã có nhiều hệ thống TTS mạnh như của IBM, Amazon, nhưng hỗ trợ tiếng Việt hiện vẫn còn hạn chế.
* **Nhận dạng ký tự quang học (OCR – Optical Character Recognition)**: Biến các tài liệu in hoặc viết tay thành văn bản số có thể chỉnh sửa và tìm kiếm được. Đây là công nghệ quan trọng trong việc số hóa tài liệu.
* **Truy hồi thông tin (Information Retrieval)**: Hệ thống tìm kiếm thông tin từ tập dữ liệu không có cấu trúc (thường là văn bản), ví dụ như các công cụ tìm kiếm Google, Yahoo, Bing hay Cốc Cốc.
* **Trích xuất thông tin (Information Extraction)**: Xác định và trích lọc các thực thể, mối quan hệ và sự kiện trong văn bản. Ứng dụng phổ biến trong các hệ thống hỏi đáp hoặc phân tích nội dung tự động.
* **Trả lời câu hỏi (Question Answering)**: Các hệ thống này có thể hiểu và trả lời câu hỏi từ người dùng dựa trên một tập dữ liệu văn bản có sẵn.
* **Tóm tắt văn bản tự động (Text Summarization)**: Rút gọn văn bản đầu vào thành một bản tóm tắt ngắn hơn nhưng vẫn truyền tải đầy đủ nội dung chính.
* **Chatbot**: Là chương trình máy tính được thiết kế để tương tác với người dùng bằng cách trò chuyện tự nhiên qua văn bản. Chatbot thường được ứng dụng trong chăm sóc khách hàng, trợ lý ảo, ... Ví dụ: ChatGPT, GROK, …
* **Dịch máy (Machine Translation)**: Cho phép chuyển đổi văn bản từ ngôn ngữ này sang ngôn ngữ khác một cách tự động. Ví dụ tiêu biểu là Google Translate.
* **Tự động kiểm tra lỗi chính tả**: Nhận diện và đề xuất sửa lỗi chính tả, ngữ pháp, ngữ nghĩa trong văn bản bằng các thuật toán NLP kết hợp mô hình ngôn ngữ.
* **Khai phá dữ liệu (Data Mining)**: Mặc dù không chỉ riêng cho NLP, nhưng khi kết hợp với xử lý văn bản, data mining giúp phát hiện mẫu và xu hướng từ khối lượng dữ liệu lớn. Đây là lĩnh vực có tiềm năng phát triển rất lớn tại Việt Nam.

# CƠ SỞ LÝ THUYẾT

Bài toán nhận diện cảm xúc thuộc nhóm các bài toán phân tích ngữ nghĩa, nơi mà mục tiêu chính là giải mã ý nghĩa ẩn sau từng câu hoặc đoạn văn để xác định sắc thái cảm xúc mà người viết muốn truyền tải. Về bản chất, đây là một dạng bài toán phân loại trong lĩnh vực khai phá dữ liệu, khi mà mỗi đoạn văn bản cần được gán vào một nhóm cảm xúc cụ thể.

Trong nghiên cứu này, chúng tôi đề xuất xây dựng một mô hình có khả năng tự động xác định cảm xúc từ văn bản đầu vào – có thể là một câu đơn, một đoạn hội thoại, hoặc một bài viết hoàn chỉnh. Đầu ra của hệ thống sẽ là nhãn cảm xúc tương ứng, chẳng hạn như tích cực, tiêu cực hoặc trung lập, tùy thuộc vào mức độ phân loại mong muốn.

Đối với những ứng dụng phổ biến như đánh giá sản phẩm, bài toán thường được đơn giản hóa bằng cách chia cảm xúc thành hai nhóm cơ bản: **tích cực** và **tiêu cực**, giúp mô hình dễ dàng học và phản hồi với hiệu quả cao hơn. Việc xác định cảm xúc này là nền tảng quan trọng trong nhiều hệ thống tương tác người – máy, đặc biệt là các nền tảng mạng xã hội, chatbot chăm sóc khách hàng, và công cụ phân tích đánh giá sản phẩm.

## Các mô hình mạng nơ-ron sử dụn trong học sâu

### Khái niệm tổng quát

Học sâu (Deep Learning) là một nhánh của học máy (Machine Learning) với điểm đặc trưng là sử dụng các mô hình có nhiều tầng xử lý để tự động trích xuất đặc trưng và học các biểu diễn dữ liệu phức tạp. Mỗi tầng trong mô hình sẽ nhận đầu vào từ tầng trước và tạo ra đầu ra cho tầng kế tiếp, giúp mô hình học được các đặc trưng từ đơn giản đến trừu tượng hơn qua từng lớp.

Một đặc điểm nổi bật của học sâu là khả năng hình thành các cấp độ biểu diễn dữ liệu theo dạng phân cấp – từ mức thấp như đặc điểm hình dạng, từ vựng, … cho tới mức cao hơn như khái niệm, ngữ nghĩa, ... Từ đó, các mô hình học sâu có thể đạt hiệu quả vượt trội trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, thị giác máy tính và nhận dạng giọng nói.

### Một số kiến trúc mạng nơ-ron phổ biến

***Mạng nơ-ron nhân tạo (ANN - Artificial Neural Network)***  
Là mô hình nền tảng mô phỏng hoạt động của các nơ-ron sinh học, ANN bao gồm nhiều lớp: lớp đầu vào, một hoặc nhiều lớp ẩn và lớp đầu ra. ANN phù hợp với nhiều tác vụ học máy như phân loại, hồi quy, …

***Mạng nơ-ron sâu (DNN - Deep Neural Network)***  
DNN là phiên bản mở rộng của ANN với nhiều lớp ẩn hơn. Việc tăng chiều sâu giúp mô hình có khả năng học các biểu diễn phức tạp hơn, đặc biệt hữu ích trong các bài toán cần trích xuất đặc trưng mạnh như phân loại ảnh hoặc văn bản.

***Mạng tích chập (CNN - Convolutional Neural Network)***  
CNN được phát triển để xử lý dữ liệu dạng lưới, chẳng hạn như hình ảnh (lưới 2 chiều). Mạng sử dụng các tầng tích chập và gộp (pooling) để tự động phát hiện các đặc trưng không gian như đường viền, góc cạnh, vùng đối tượng… CNN đã đạt nhiều thành tựu trong thị giác máy tính, và gần đây cũng được ứng dụng hiệu quả trong NLP (xử lý ngôn ngữ).

***Mạng bộ nhớ dài ngắn hạn (LSTM - Long Short-Term Memory)***  
Được giới thiệu bởi Hochreiter và Schmidhuber năm 1997, LSTM là một biến thể của mạng nơ-ron hồi tiếp (RNN - Recurrent Neural Network), được thiết kế để giải quyết vấn đề quên thông tin trong các chuỗi dài. LSTM đặc biệt hiệu quả trong các bài toán liên quan đến chuỗi thời gian và ngữ cảnh dài như dịch máy, sinh văn bản, hoặc nhận diện chữ viết tay.

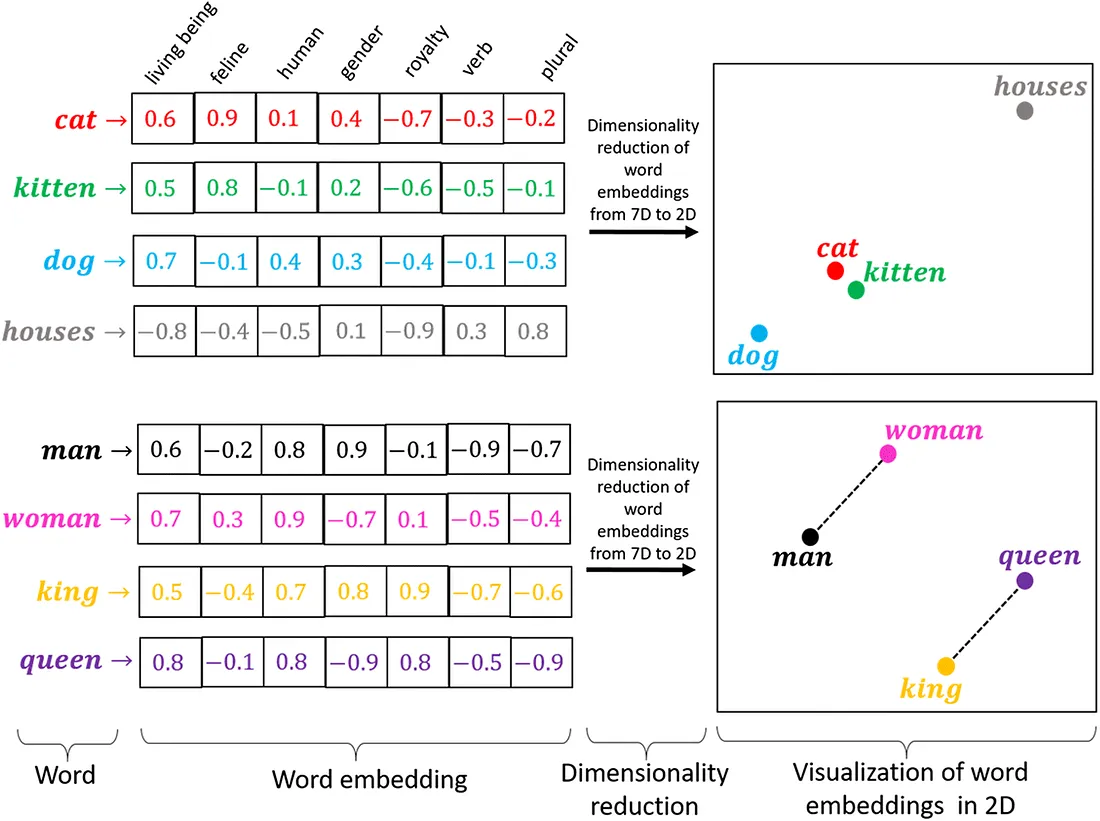
Ngoài ra, một số kỹ thuật còn sử dụng huấn luyện không giám sát ban đầu để khởi tạo trọng số, sau đó tinh chỉnh lại bằng học có giám sát (supervised learning) để đạt hiệu quả tốt hơn trong phân loại hoặc gán nhãn dữ liệu.

## Biểu diễn văn bản bằng Word2Vec

### Khái quát chung

Word2Vec là một kỹ thuật biểu diễn từ vựng trong không gian vector sao cho các mối quan hệ ngữ nghĩa và cú pháp giữa các từ được bảo toàn ở mức độ nhất định. Thay vì gán cho mỗi từ một chỉ số duy nhất như trong one-hot encoding, Word2Vec ánh xạ mỗi từ thành một vector có giá trị thực trong không gian nhiều chiều, phản ánh cách từ đó tương quan với các từ khác trong ngữ cảnh.

Thông qua việc học từ dữ liệu huấn luyện, mỗi từ sẽ có một biểu diễn phân bố (distributed representation), nghĩa là không còn gắn liền với một chiều cố định trong vector, mà được định nghĩa bởi nhiều thành phần – và mỗi thành phần lại đồng thời đóng góp vào biểu diễn của nhiều từ khác. Nhờ vậy, phương pháp này mang lại sự linh hoạt và hiệu quả cao trong việc xử lý ngôn ngữ tự nhiên.



Hình .1 Cách biểu diễn các từ trên Word2Vec

### Nguyên lý hoạt động

Quá trình tạo ra các vector từ thông qua Word2Vec bao gồm hai giai đoạn chính:

1. **Xây dựng từ điển (Vocabulary)**: Trích xuất tất cả từ xuất hiện trong tập văn bản đầu vào (corpus), sau khi xử lý sơ bộ như loại bỏ dấu câu, chuyển chữ thường …
2. **Huấn luyện mô hình**: Dựa trên sự xuất hiện đồng thời của các từ trong cùng ngữ cảnh, mô hình sẽ học cách ánh xạ mỗi từ thành một vector sao cho phản ánh được ngữ nghĩa tương đối giữa các từ trong ngữ cảnh đó.

Kết quả sau quá trình huấn luyện là một tập các vector từ (word vectors), có thể lưu trữ dưới định dạng văn bản hoặc nhị phân. Những vector này có thể được sử dụng trực tiếp trong các bài toán học máy như phân loại văn bản, phân tích cảm xúc, phân cụm, …

### Các kiến trúc mô hình

Hai kiến trúc phổ biến được sử dụng để huấn luyện Word2Vec bao gồm:

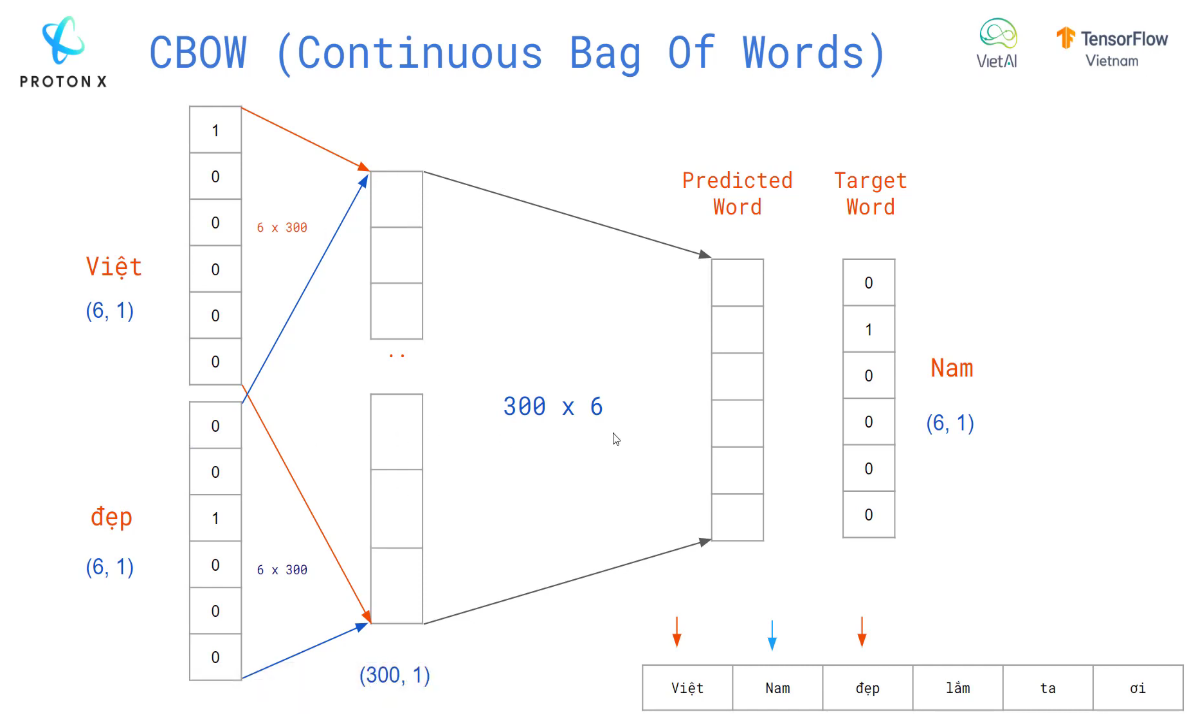
**1. Continuous Bag of Words (CBOW)**

Mô hình CBOW học bằng cách dự đoán một từ trung tâm dựa vào các từ xung quanh nó trong một phạm vi nhất định (window size). Cụ thể, mô hình lấy các từ ở hai bên (trái và phải) làm ngữ cảnh, sau đó dự đoán từ cần tìm nằm giữa.

Đặc điểm của CBOW:

* Đầu vào: các từ ngữ cảnh xung quanh.
* Đầu ra: từ trung tâm cần dự đoán.
* Thứ tự các từ trong ngữ cảnh không ảnh hưởng đến kết quả dự đoán.
* Phù hợp với tập dữ liệu lớn, thường cho kết quả ổn định với tốc độ huấn luyện nhanh.

**2. Skip-Gram**



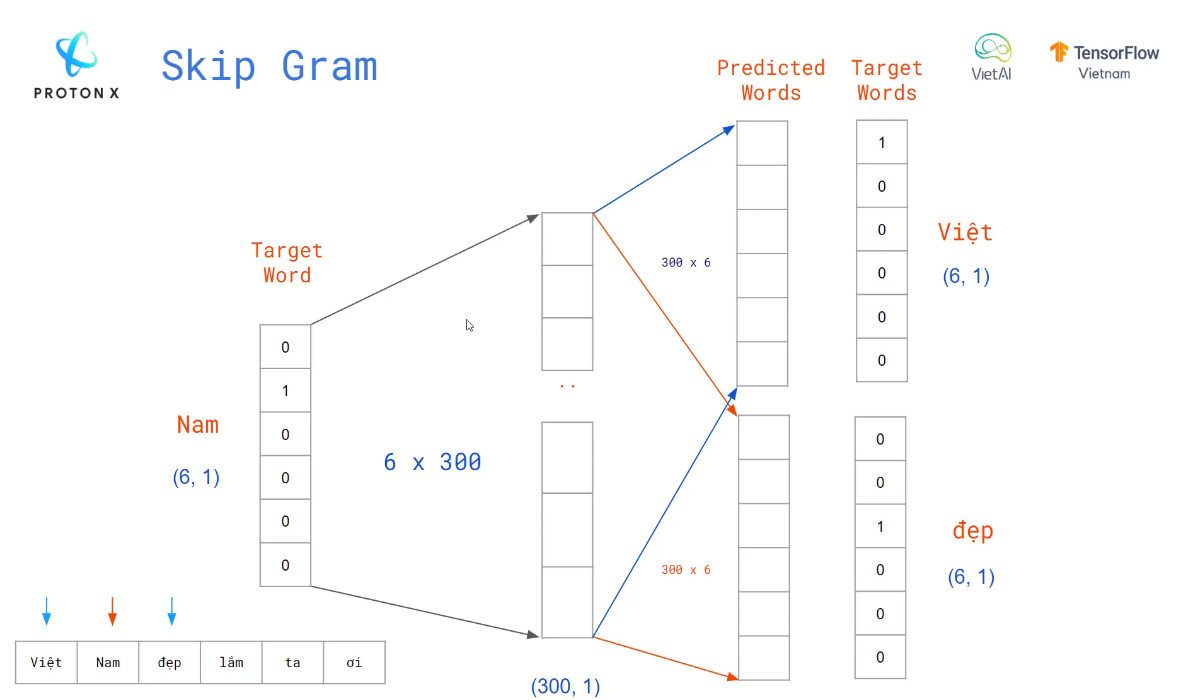
Hình .2 Mô hình CBOW

Trái ngược với CBOW, mô hình Skip-Gram dự đoán các từ ngữ cảnh dựa trên một từ trung tâm. Đây là phương pháp mạnh mẽ, đặc biệt hiệu quả khi làm việc với các từ hiếm, bởi nó cố gắng tận dụng từ trung tâm để học các mối quan hệ với nhiều từ khác nhau trong ngữ cảnh.

Đặc điểm của Skip-Gram:

* Đầu vào: từ trung tâm.
* Đầu ra: các từ xung quanh trong phạm vi ngữ cảnh.
* Có thể điều chỉnh phạm vi từ (context window) để cân bằng giữa chất lượng biểu diễn và độ phức tạp tính toán.
* Skip-Gram cho phép gán trọng số nhỏ hơn cho những từ ngữ cảnh ở xa nhằm giảm ảnh hưởng của các mối liên hệ yếu.

Mô hình Skip-Gram được tối ưu để tính toán hiệu quả, không sử dụng phép nhân ma trận dày đặc, cho phép huấn luyện nhanh trên tập dữ liệu rất lớn.



Hình .3 Mô hình Skip-Gram

### Thuật toán tối ưu

Để cải thiện hiệu quả huấn luyện, Word2Vec sử dụng hai kỹ thuật tối ưu chính:

* **Hierarchical Softmax**: Từ vựng được tổ chức thành một cây nhị phân, trong đó mỗi từ là một lá của cây. Khi cần dự đoán một từ, mô hình chỉ cần tính toán dọc theo đường đi từ gốc đến lá tương ứng, giúp giảm chi phí tính toán từ O(V) xuống O(log V), với V là số lượng từ trong từ điển.
* **Negative Sampling**: Thay vì cập nhật trọng số cho toàn bộ từ điển, phương pháp này chỉ cập nhật một số lượng nhỏ các từ trong mỗi vòng lặp huấn luyện – bao gồm một từ đúng (positive sample) và một số từ sai (negative samples) được chọn ngẫu nhiên. Điều này giúp rút ngắn thời gian huấn luyện đáng kể mà vẫn đảm bảo chất lượng vector học được.

## Các mô hình nhận diện cảm xúc trong văn bản

### Tiếp cận phân tích cảm xúc từ xử lý ngôn ngữ tự nhiên

Các phản hồi, đánh giá hoặc bình luận do người dùng để lại trên các nền tảng trực tuyến, như website bán hàng hay mạng xã hội, là một dạng ngôn ngữ tự nhiên được tạo ra một cách tự phát. Để xử lý dữ liệu này trong phân tích cảm xúc, bước đầu tiên là thu thập tập dữ liệu chứa các nội dung nhận xét, bình luận của người dùng sau khi đã trải nghiệm sản phẩm hoặc dịch vụ.

Sau khi thu thập, dữ liệu cần được **làm sạch**: loại bỏ ký tự đặc biệt, các thành phần không mang ý nghĩa, xử lý lỗi chính tả, chuẩn hóa định dạng ngữ pháp, loại bỏ từ dừng… Đây là giai đoạn tiền xử lý giúp chuẩn bị dữ liệu đầu vào có chất lượng tốt cho mô hình.

Tiếp đến, người nghiên cứu tiến hành khảo sát tổng quan tập dữ liệu – bao gồm xác định kích thước tập dữ liệu, đặc tính phân bố cảm xúc, từ vựng thường gặp, ... Từ đó, chọn lựa các đặc trưng phù hợp để đưa vào mô hình phân tích. Việc xác định đúng các chiều dữ liệu (feature selection) là bước then chốt, bởi các đặc trưng càng mang tính đại diện cao thì mô hình càng dễ đưa ra kết quả chính xác.

Cuối cùng, kết quả phân tích sẽ được đánh giá qua các chỉ số như độ chính xác, F1-score, từ đó làm cơ sở cho việc triển khai ứng dụng vào thực tế hoặc điều chỉnh mô hình để cải thiện hiệu quả.

### Tiếp cận phân tích cảm xúc bằng phương pháp học máy

Phân tích cảm xúc có thể được hiểu là quá trình đánh giá và nhận diện thái độ, quan điểm, hay cảm xúc của người dùng thông qua ngôn ngữ trong văn bản. Điều này đã trở thành chủ đề trọng tâm trong nhiều nghiên cứu liên quan đến mạng xã hội và dữ liệu người dùng (Liu, 2012; Pang & Lee, 2008).

Trong bối cảnh số lượng dữ liệu văn bản ngày càng lớn, đặc biệt từ các nền tảng số, việc áp dụng **học máy** (machine learning) – một nhánh quan trọng của trí tuệ nhân tạo – để tự động hóa phân tích cảm xúc là hướng tiếp cận phổ biến.

Học máy được chia thành bốn nhóm chính:

* **Học có giám sát** (Supervised Learning): Hệ thống học từ dữ liệu đã được gán nhãn để dự đoán nhãn cho dữ liệu mới. Bao gồm các bài toán như phân loại (classification) và hồi quy (regression).
* **Học không giám sát** (Unsupervised Learning): Dữ liệu đầu vào không có nhãn. Hệ thống tự phát hiện cấu trúc hoặc mẫu ẩn thông qua phân cụm (clustering) hay phát hiện luật (association rule).
* **Học bán giám sát** (Semi-supervised Learning): Kết hợp giữa hai phương pháp trên, một phần dữ liệu được gán nhãn, phần còn lại không có nhãn.
* **Học tăng cường** (Reinforcement Learning): Mô hình học qua quá trình thử - sai, nhận phần thưởng hoặc hình phạt để điều chỉnh hành vi.

Trong nghiên cứu này, phương pháp **học có giám sát** được lựa chọn để xây dựng hệ thống nhận diện cảm xúc từ các phản hồi của người dùng.

### Kiến trúc tổng quát của mô hình

Quá trình nghiên cứu được thực hiện theo các bước cơ bản sau:

1. **Thu thập dữ liệu**: Dữ liệu văn bản được lấy từ nguồn thực tế, cụ thể là từ các bình luận, đánh giá trên trang web *Foody.vn*. Đây là nguồn dữ liệu chứa nhiều phản hồi thực tế, phù hợp với bài toán phân tích cảm xúc.
2. **Tiền xử lý và gán nhãn**: Trong nghiên cứu này, dữ liệu văn bản được cung cấp dưới dạng đã được tiền xử lý sơ bộ, bao gồm các bước làm sạch như: loại bỏ ký tự đặc biệt, xóa các từ không mang giá trị ngữ nghĩa (stop words), chuẩn hóa dấu câu và định dạng văn bản.

* Để chuẩn bị cho bước vector hóa và huấn luyện mô hình, tôi thực hiện thêm một bước **chuẩn hóa từ vựng**, cụ thể là thay thế khoảng trắng giữa các từ trong cụm bằng dấu gạch dưới (\_). Ví dụ: cụm từ “tiếng Việt” được chuyển thành “tiếng\_Việt”. Việc này giúp mô hình phân tích ngữ nghĩa chính xác hơn, đặc biệt trong các kỹ thuật trích xuất đặc trưng như Word2Vec hoặc TF-IDF, nơi các từ ghép cần được giữ nguyên làm một thực thể thống nhất.
* Sau khi hoàn tất, các câu văn trong tập dữ liệu sẽ có định dạng rõ ràng, nhất quán và thuận lợi cho việc tách từ và huấn luyện mô hình phân loại cảm xúc. Cuối cùng, từng file liệu được gán nhãn cảm xúc tương ứng (ví dụ: “tích cực” hoặc “tiêu cực”), phục vụ cho bài toán học có giám sát.

1. **Chia dữ liệu huấn luyện và kiểm thử**:
   * **Tập huấn luyện**: Được sử dụng để đào tạo mô hình học máy. Các văn bản sau tiền xử lý được biến đổi thành các vector đặc trưng thông qua các phương pháp như TF-IDF, Word2Vec... Mô hình học từ mối quan hệ giữa đặc trưng và nhãn để tìm ra quy luật phân loại.
   * **Tập kiểm thử**: Sau khi mô hình học xong, tập dữ liệu kiểm thử được sử dụng để đánh giá khả năng dự đoán của mô hình với dữ liệu chưa từng thấy. Dữ liệu trong tập này cũng đã được tiền xử lý tiếp đó được chuyển về vector đặc trưng như tập huấn luyện, sau đó đưa vào mô hình để dự đoán nhãn cảm xúc.

Việc đánh giá hiệu suất mô hình được thực hiện qua các chỉ số đo lường chất lượng như: độ chính xác (accuracy), độ thu hồi (recall), độ chính xác (precision), F1-score...

# NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT

## Tiền xử lý dữ liệu văn bản

Trong lĩnh vực ngôn ngữ học, “ngôn ngữ tự nhiên” được hiểu là các loại ngôn ngữ hình thành một cách tự nhiên trong quá trình tiến hóa và giao tiếp của loài người, không phải là kết quả của sự lập trình hay thiết kế có chủ đích. Đây là phương tiện giao tiếp cơ bản giữa con người với nhau và có thể tồn tại dưới nhiều hình thức khác nhau như lời nói, chữ viết, ngôn ngữ ký hiệu hay các dạng cảm nhận xúc giác.

Trong bài toán xử lý ngôn ngữ tự nhiên tiếng Việt, trước khi dữ liệu được đưa vào mô hình học máy, một bước quan trọng không thể bỏ qua là **tách từ** và **chuẩn hóa từ ngữ**. Đây là khâu tiền xử lý đóng vai trò nền tảng, giúp mô hình hiểu đúng ngữ nghĩa văn bản, từ đó nâng cao chất lượng của các tác vụ như phân tích cảm xúc, phân loại, tóm tắt hay dịch tự động.

Trong nghiên cứu này, dữ liệu được thu thập từ nguồn đã qua bước làm sạch sơ bộ (xóa HTML, icon, ký tự đặc biệt, dòng trống...). Việc xử lý tập trung chủ yếu vào chuẩn hóa từ ngữ và chính tả, thông qua từ điển ánh xạ các từ viết tắt phổ biến về dạng chuẩn. Việc này được thực hiện bán tự động bằng script Python kết hợp với xử lý chuỗi.

### Tách từ

Tách từ (word segmentation) là một bước không thể thiếu trong tiền xử lý ngữ liệu tiếng Việt. Khác với tiếng Anh – nơi dấu cách được dùng để phân biệt ranh giới từ – thì trong tiếng Việt, dấu cách lại chỉ đóng vai trò phân tách các **âm tiết**, chứ không phân định từ hoàn chỉnh. Điều này khiến cho việc tách từ trong tiếng Việt trở nên phức tạp và là một thách thức trong các bài toán xử lý ngôn ngữ tự nhiên.

Ví dụ, cụm từ “đất nước” bao gồm hai âm tiết “đất” và “nước”, mỗi âm tiết đều có thể tồn tại độc lập với nghĩa riêng. Tuy nhiên, khi ghép lại, chúng tạo thành một từ mang ý nghĩa khác biệt. Nếu không tách đúng cụm “đất nước”, mô hình học máy có thể hiểu nhầm hoặc xử lý sai ngữ cảnh.

**Mục tiêu và vai trò**

Tách từ giúp hệ thống:

* Hiểu rõ ranh giới từ vựng.
* Loại bỏ nhập nhằng về mặt ngữ nghĩa.
* Tăng độ chính xác khi chuyển văn bản thành các đặc trưng đầu vào cho mô hình học máy.

Một ví dụ minh họa rõ nét cho vấn đề nhập nhằng ngữ nghĩa nếu tách sai:

* “Ăn cơm không được uống rượu.” có thể tách thành:
  + Ăn / cơm / không / được / uống / rượu.
  + Ăn / cơm không / được / uống / rượu.

Tùy vào cách tách, nghĩa của câu sẽ thay đổi hoàn toàn.

**Quy tắc biểu diễn sau khi tách**

Sau khi tách, các từ ghép sẽ được nối bằng ký tự gạch dưới (\_) để biểu diễn như một cụm từ thống nhất. Ví dụ:

“ẩm thực Việt Nam” → ẩm\_thực Việt\_Nam

Đây là quy ước phổ biến trong các hệ thống xử lý ngôn ngữ tự nhiên tiếng Việt, nhằm giữ được mạch ngữ nghĩa cho câu.

**Các phương pháp tách từ phổ biến**

Hiện nay có bốn hướng tiếp cận chính trong việc tách từ tiếng Việt:

1. **Dựa vào từ điển** – So khớp các chuỗi từ trong văn bản với từ điển từ vựng có sẵn.
2. **Dựa vào thống kê** – Sử dụng tần suất xuất hiện của các cụm từ trong tập dữ liệu để xác định ranh giới từ.
3. **Kết hợp từ điển và thống kê** – Phương pháp kết hợp giúp tận dụng cả ưu điểm của từ điển và tính thực tiễn từ dữ liệu.
4. **Dựa trên ký tự (n-gram)** – Tách câu thành các chuỗi ký tự dài cố định như unigram, bigram... Dễ triển khai nhưng hạn chế về mặt ngữ nghĩa.

Trong nghiên cứu này, quá trình tách từ được thực hiện thông qua công cụ Tokenizer chuyên biệt cho tiếng Việt cụ thể là tool VnTokenizer. Việc sử dụng công cụ giúp tự động xác định ranh giới từ, tăng độ chính xác và nhất quán cho toàn bộ tập dữ liệu.

### Chuẩn hóa từ ngữ và chính tả

Bên cạnh tách từ, bước **chuẩn hóa từ ngữ và chính tả** cũng giữ vai trò quan trọng trong việc làm sạch và làm giàu dữ liệu đầu vào. Trong môi trường mạng xã hội, blog, hoặc các nền tảng bình luận, người dùng thường xuyên sử dụng **viết tắt**, **teencode**, hoặc **sai chính tả** – nếu không xử lý, mô hình học máy sẽ xem đây là các từ độc lập, gây nhiễu và làm giảm độ chính xác khi huấn luyện và dự đoán.

**Tại sao cần chuẩn hóa?**

Cùng một từ ngữ có thể được viết dưới nhiều dạng khác nhau, chẳng hạn:

* không có thể được viết thành: ko, k, kh
* với → vs
* rồi → r, rùi, r
* cũng → cx
* như thế nào → ntn
* gì → j

Nếu giữ nguyên các dạng biểu diễn như vậy, dữ liệu sẽ bị phân mảnh, mô hình khó học được các đặc trưng nhất quán. Việc chuẩn hóa đưa mọi từ viết tắt, sai chính tả, hoặc biểu hiện không chính quy về một **dạng chuẩn thống nhất**, giúp:

* Giảm số lượng từ vựng không cần thiết.
* Tránh hiểu nhầm ngữ nghĩa.
* Tối ưu dung lượng lưu trữ và tốc độ xử lý.

Sau đó văn bản được quét và thay thế từ sai thành dạng chuẩn. Điều này giúp dữ liệu đầu vào nhất quán, giúp mô hình Word2Vec hoặc các vector hóa khác học biểu diễn từ một cách hiệu quả hơn.

## Vector hóa văn bản

Trong các hệ thống phân tích cảm xúc sử dụng mô hình học máy, một yêu cầu tiên quyết là dữ liệu văn bản phải được chuyển đổi thành dạng số để máy tính có thể xử lý. Bởi lẽ, các thuật toán không thể làm việc trực tiếp với từ ngữ hay câu chữ như con người, mà chỉ hiểu và tính toán trên các con số.

Quá trình ánh xạ các đơn vị từ vựng (như từ, cụm từ) sang không gian số được gọi là **vector hóa văn bản**. Đây là bước quan trọng giúp biến đổi văn bản thành các vector đặc trưng trong không gian nhiều chiều, phản ánh phần nào ngữ nghĩa và mối quan hệ giữa các từ.

Trong nghiên cứu này, chúng tôi sử dụng hai hướng tiếp cận chính:

* Các phương pháp **embedding cổ điển** (như Bag of Words, TF-IDF, Distributional).
* Các phương pháp **embedding hiện đại** dựa trên mạng nơ-ron (Word2Vec, GloVe, FastText).

### Các phương pháp biểu diễn văn bản cổ điển (Classic Word Embedding)

**• *Bag of Words (BoW)***

Phương pháp Bag of Words (túi từ) là cách tiếp cận đơn giản nhưng hiệu quả ở mức cơ bản. Mỗi văn bản được biểu diễn bằng một vector có số chiều tương ứng với kích thước của bộ từ vựng. Nếu một từ xuất hiện trong văn bản, vị trí tương ứng trong vector sẽ được gán giá trị 1, nếu không thì gán 0.

Tuy nhiên, mô hình này không tính đến tần suất xuất hiện hoặc ngữ cảnh sử dụng từ, nên thường không phản ánh được ý nghĩa đầy đủ.



Hình 3.1 Mô hình biểu diễn Bag of Words

**• *TF-IDF (Term Frequency - Inverse Document Frequency)***

TF-IDF là phương pháp cải tiến từ BoW, giúp đánh giá **tầm quan trọng** của một từ trong một văn bản so với toàn bộ tập tài liệu. Cách làm này không chỉ quan tâm đến sự xuất hiện, mà còn xét đến sự hiếm gặp của từ đó trong toàn bộ tập dữ liệu – từ càng hiếm nhưng xuất hiện nhiều trong một văn bản thì càng có trọng số cao.

Công thức tính:

* TF (Term frequency): Tần suất xuất hiện của một từ trong một văn bản.

Công thức:

Trong đó:

* + : số lần từ *i* xuất hiện trong văn bản
  + : tổng số từ trong văn bản
* IDF (Invert Document Frequency): Hay tần số văn bản nghịch đảo, được dùng để đánh giá tầm quan trọng của từ trong văn bản.

Công thức:

Trong đó:

* + : tổng số văn bản có trong tập dữ liệu (ví dụ: một thư mục chứa 500 file văn bản có định dạng .txt, thì D = 500)
  + : số văn bản có chứa từ
* TF-IDF (Term frequency - Invert Document Frequency): Sự kết hợp của tần số từ tf và tần số văn bản nghịch đảo idf.

Công thức:

**• *Distributional Embedding***

**Distributional Embedding** là một cách biểu diễn từ (word) trong không gian số (vector space) dựa trên giả thuyết:

**“You shall know a word by the company it keeps.”** – J.R. Firth (Tạm dịch: “Bạn có thể hiểu một từ qua những từ xung quanh nó.”)

Đây là phương pháp biểu diễn từ dựa trên **mối liên hệ phân bố** với các từ xung quanh. Thay vì chỉ đơn thuần đếm từ, Distributional Embedding xây dựng một ma trận biểu diễn sự đồng xuất hiện giữa các từ trong ngữ cảnh. Mỗi hàng trong ma trận tương ứng với một từ, còn các cột thể hiện tần suất đồng xuất hiện với các từ khác trong một phạm vi ngữ cảnh cố định.

Phương pháp này phản ánh rõ hơn mối quan hệ ngữ nghĩa giữa các từ, nhưng ma trận biểu diễn thường có kích thước lớn và gây tốn tài nguyên tính toán.



Hình 3.2 Ma trận biểu diễn theo phương pháp Distributional Embedding

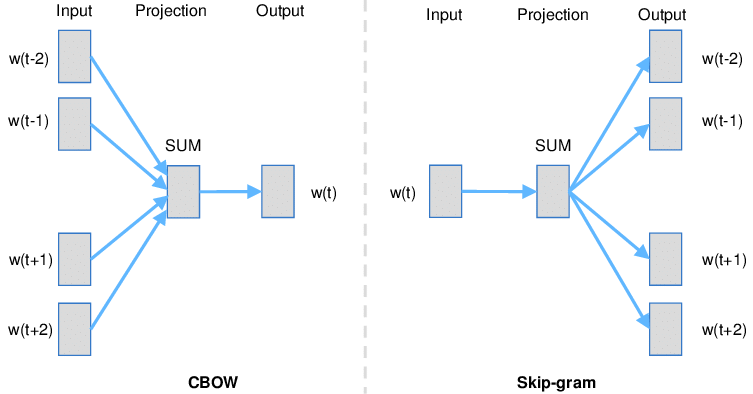
### Các phương pháp biểu diễn văn bản hiện đại

**• *Word2Vec***

Word2Vec là một trong những phương pháp embedding hiện đại phổ biến nhất hiện nay. Thay vì xây dựng ma trận đếm hay thống kê thủ công, Word2Vec sử dụng một mạng nơ-ron đơn giản để học biểu diễn vector cho các từ dựa vào ngữ cảnh xuất hiện.

Hai kiến trúc chính:

* **CBOW (Continuous Bag of Words)**: Dùng các từ xung quanh để dự đoán từ trung tâm.
* **Skip-gram**: Dùng một từ để dự đoán các từ xung quanh nó trong một cửa sổ ngữ cảnh.



Hình 3.3 Mô hình CBOW và Skip-Gram

Skip-gram có ưu điểm học được tốt cả với những từ ít xuất hiện nhưng huấn luyện tốn thời gian hơn so với CBOW.

**• *FastText***

FastText là phiên bản cải tiến của Word2Vec, được phát triển bởi Facebook AI. Khác với Word2Vec – học embedding cho toàn bộ từ – thì FastText chia từ thành các đơn vị nhỏ hơn gọi là n-gram ký tự (ví dụ: ng, gra, ram). Việc học biểu diễn dựa trên các n-gram giúp mô hình xử lý tốt hơn các từ hiếm, từ chưa thấy hoặc từ viết sai chính tả.

Điểm mạnh của FastText là giúp tăng khả năng khái quát và cải thiện đáng kể độ chính xác của các mô hình xử lý ngôn ngữ tự nhiên, đặc biệt là tiếng Việt – nơi có nhiều biến thể từ và lỗi chính tả thường gặp.

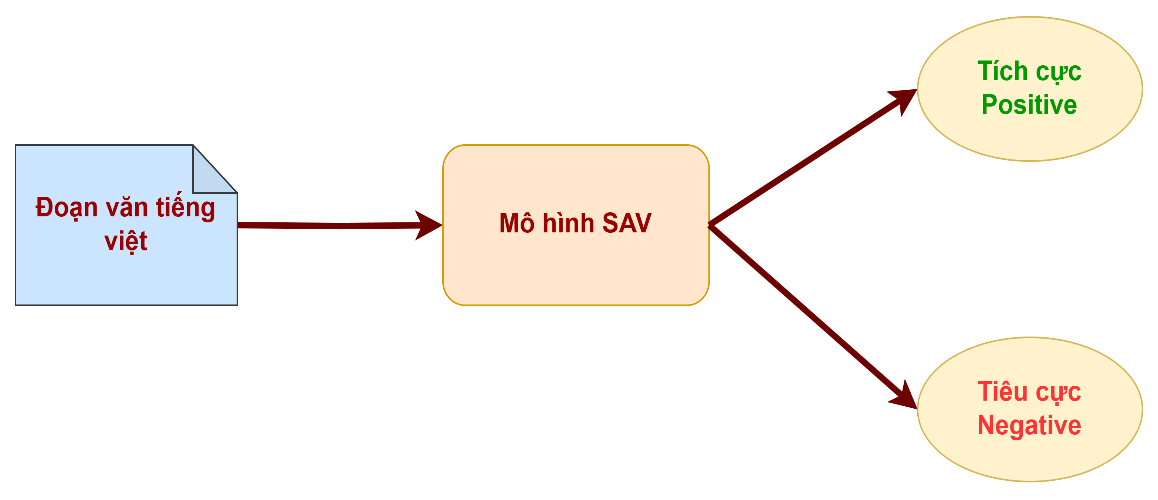
## Mô hình nhận diện cảm xúc sử dụng học sâu và học máy

Để giải quyết bài toán phân loại cảm xúc trong văn bản tiếng Việt, đồ án này áp dụng các phương pháp học sâu và học máy cổ điển kết hợp. Cụ thể, chúng tôi sử dụng mô hình mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) với lớp LSTM (Long Short-Term Memory) nhằm xử lý chuỗi văn bản có tính tuần tự, đồng thời kết hợp với các mô hình học máy giám sát như Naive Bayes và SVM (Support Vector Machine) để huấn luyện các mô hình truyền thống với các đặc trưng văn bản đã được vector hóa.

Quá trình huấn luyện các mô hình này sử dụng hai phương pháp biểu diễn văn bản chính là **TF-IDF** và **Word2Vec**. Với Word2Vec, chúng tôi sử dụng kiến trúc **Continuous Bag of Words (CBOW)** để học biểu diễn từ vựng dựa trên ngữ cảnh xung quanh, trong khi TF-IDF giúp gán trọng số cho từng từ theo mức độ quan trọng trong toàn bộ tập dữ liệu. Sự kết hợp này cho phép mô hình vừa tận dụng ngữ nghĩa cục bộ từ ngữ cảnh, vừa lọc ra các từ có ý nghĩa nổi bật nhất trong phân tích cảm xúc.

Để mô hình đạt được hiệu quả tốt, dữ liệu đầu vào cần có số lượng đủ lớn và đã được gán nhãn rõ ràng theo từng loại cảm xúc (ví dụ: tích cực hoặc tiêu cực). Một phần dữ liệu được sử dụng để huấn luyện, phần còn lại dùng để đánh giá độ chính xác của các mô hình. Mô hình học sâu được xây dựng bằng Keras, với lớp Embedding sử dụng trọng số từ Word2Vec tiền huấn luyện, kết hợp với lớp LSTM để học các mối quan hệ ngữ nghĩa và ngữ cảnh theo chuỗi.

Ngoài ra, các mô hình truyền thống như Naive Bayes và SVM cũng được triển khai song song để so sánh hiệu quả với mô hình học sâu. Cả ba mô hình được huấn luyện trên cùng tập dữ liệu đầu vào và đánh giá dựa trên các chỉ số như độ chính xác, độ nhạy, và độ đặc hiệu.



Hình 3.4 Mô hình SAV

# THỰC NGHIỆM

## Xây dựng dữ liệu

Trong lĩnh vực ngôn ngữ học, “ngôn ngữ tự nhiên” được hiểu là các loại ngôn ngữ hình thành một cách tự nhiên trong quá trình tiến hóa và giao tiếp của loài người, không phải là kết quả của sự lập trình hay thiết kế có chủ đích. Đây là phương tiện giao tiếp cơ bản giữa con người với nhau và có thể tồn tại dưới nhiều hình thức khác nhau như lời nói, chữ viết, ngôn ngữ ký hiệu hay các dạng cảm nhận xúc giác.

### Bối cảnh và vai trò của dữ liệu huấn luyện

Trong lĩnh vực học sâu (deep learning), dữ liệu không chỉ là yếu tố đầu vào – mà còn là nền móng cốt lõi quyết định mức độ thành công của mô hình. Các mô hình mạng nơ-ron sâu thường chứa hàng trăm nghìn, thậm chí hàng triệu tham số cần được tối ưu hóa qua nhiều vòng lặp. Do đó, yêu cầu về một tập dữ liệu huấn luyện có kích thước đủ lớn, chất lượng tốt và mang tính đa dạng cao là điều tất yếu.

Tuy nhiên, điều này lại đặt ra không ít thách thức cho các cá nhân hoặc nhóm nghiên cứu không trực thuộc các tổ chức có sẵn hệ sinh thái dữ liệu lớn (Big Data). Việc tự thu thập, gán nhãn và xử lý dữ liệu là công việc mất nhiều thời gian, công sức và tài nguyên. Theo kinh nghiệm của chúng tôi, có thể khẳng định rằng hơn 70–80% thời lượng nghiên cứu cho một dự án học sâu thường được dành cho khâu xử lý và chuẩn bị dữ liệu. Điều đó càng nhấn mạnh vai trò thiết yếu của việc xây dựng một tập dữ liệu huấn luyện phù hợp và có tính ứng dụng cao trong thực tiễn.

### Bộ dữ liệu nhận xét được gán nhãn cảm xúc tích cực, tiêu cực

Dựa trên định hướng ứng dụng vào bài toán phân tích cảm xúc trong ngôn ngữ tiếng Việt, tôi đã chủ động tìm kiếm một bộ dữ liệu chứa các phản hồi, đánh giá thực tế của người tiêu dùng từ các nền tảng thương mại điện tử và bộ dữ liệu này được lấy từ các bình luận của trang web foody.vn. Đây là một trong những môi trường thực tế chứa lượng lớn bình luận, mang đậm tính chủ quan và cảm xúc của người dùng – một nguồn ngữ liệu lý tưởng để khai thác cho mục tiêu nghiên cứu.

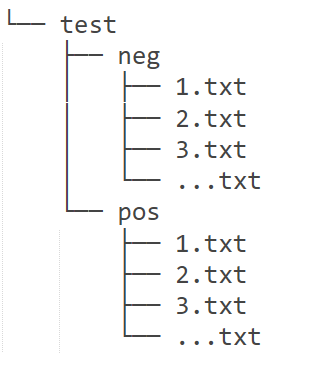
Việc gán nhãn dữ liệu được tiến hành theo cách xem xét dựa trên **mức độ hài lòng thông qua số điểm trung bình của mỗi bình luận**:

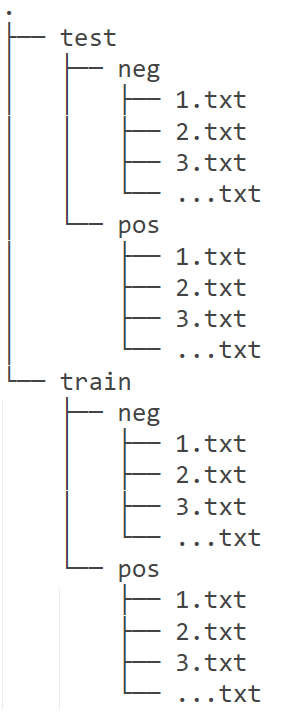
* Những phản hồi có điểm đánh giá từ **8.5 đến 10 điểm** được xếp vào nhóm **tích cực**.
* Những phản hồi có điểm đánh giá từ **0.1 đến 5.7 điểm** được xếp vào nhóm **tiêu cực**.

Sau khi thu thập, dữ liệu được chuẩn hóa và lưu trữ dưới dạng tệp .txt với:

* Nội dung phản hồi của khách hàng là nội dung của từng file.
* Còn nhãn dựa trên thư mục chứa file.

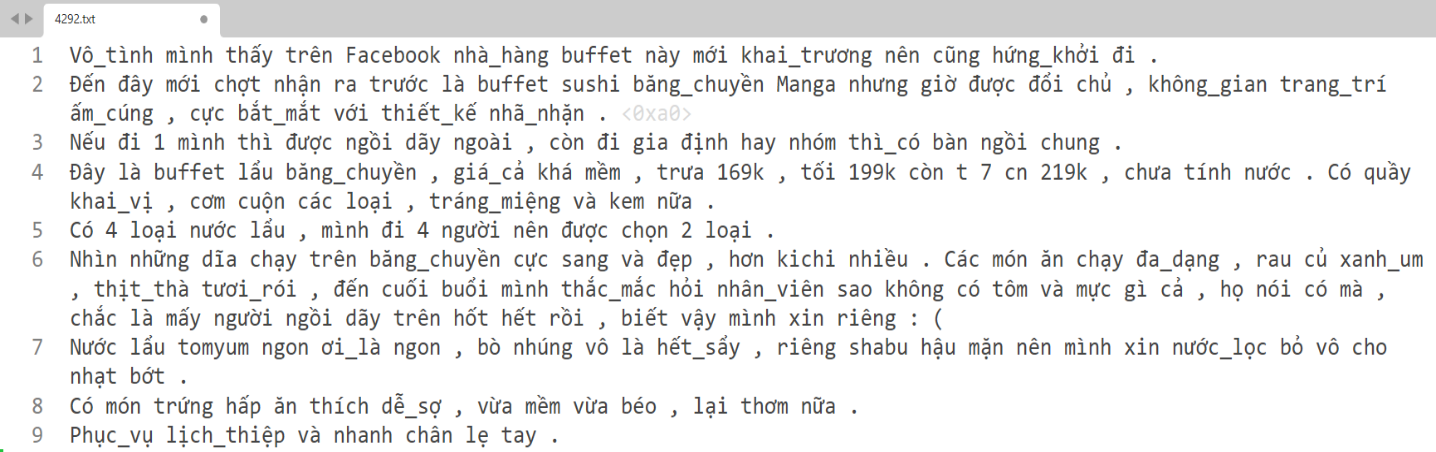
Cụ thể:

* Cấu trúc thư mục data\_train (bên trái) và data\_test (bên phải):



Hình 4.1 Cấu trúc thư mục data\_train, và data\_test

* Mô tả dữ liệu của một file:



Hình 4.2 Mô tả dữ liệu của một file

Tổng số lượng mẫu trong của cả data\_train và data\_test là **50.000 mẫu** dữ liệu bình luận:

* **25.000 mẫu tích cực**
* **25.000 mẫu tiêu cực**

Phân bố chi tiết giữa tập huấn luyện và tập kiểm tra được thể hiện trong bảng dưới đây:

Bảng 4.1 Cấu trúc bộ dữ liệu data\_train

|  |  |  |
| --- | --- | --- |
| **Tập dữ liệu** | **Tích cực** | **Tiêu cực** |
| **Huấn luyện (train)** | 15.000 | 15.000 |
| **Kiểm tra (validate)** | 5.000 | 5.000 |

Bảng 4.2 Cấu trúc bộ dữ liệu data\_test

|  |  |  |
| --- | --- | --- |
| **Tập dữ liệu** | **Tích cực** | **Tiêu cực** |
| **Kiểm tra (test)** | 5.000 | 5.000 |

## Tiền xử lý dữ liệu

Trong bất kỳ hệ thống học máy nào, đặc biệt là các mô hình xử lý ngôn ngữ tự nhiên (NLP), bước tiền xử lý dữ liệu đầu vào đóng vai trò then chốt trong việc nâng cao hiệu quả và độ chính xác của mô hình. Dữ liệu thô thu thập được thường chứa nhiều thông tin dư thừa, không đồng nhất và có thể gây nhiễu nếu được sử dụng trực tiếp để huấn luyện. Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện.

Đối với bài toán nhận diện cảm xúc trong văn bản tiếng Việt mà đồ án này hướng tới, dữ liệu đầu vào là các đánh giá, nhận xét thực tế của người dùng về sản phẩm trên các nền tảng thương mại điện tử. Những phản hồi này thường không tuân theo quy chuẩn văn phong chính thống, đôi khi còn chứa ký hiệu đặc biệt, lỗi chính tả hoặc văn bản không liên quan. Để khắc phục điều đó, quy trình tiền xử lý dữ liệu được xây dựng theo các bước sau:

* **Tách từ**: Đây là bước đặc biệt quan trọng trong xử lý tiếng Việt, do tiếng Việt sử dụng khoảng trắng để phân chia âm tiết chứ không phải ranh giới từ hoàn chỉnh. Việc xác định chính xác ranh giới từ giúp mô hình tránh hiểu sai ngữ nghĩa. Trong nghiên cứu này, chúng tôi sử dụng công cụ VnTokenizer để tách từ tự động, đồng thời giữ nguyên quy tắc biểu diễn từ ghép bằng dấu gạch dưới, ví dụ: "ẩm thực Việt Nam" sẽ được biểu diễn thành "ẩm\_thực Việt\_Nam".
* **Chuẩn hóa văn bản**: Bằng cách ánh xạ các từ viết tắt, teencode hoặc từ sai chính tả về dạng chuẩn, ví dụ: "không" → "ko", "rồi" → "r", "gì" → "j", v.v. Công đoạn này giúp giảm nhiễu và đồng nhất biểu diễn từ trong tập dữ liệu.

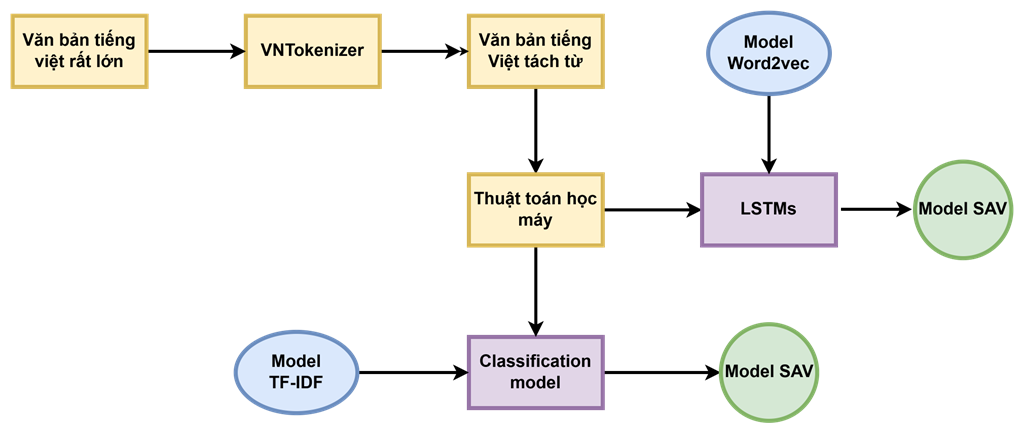
Sau khi hoàn tất bước làm sạch văn bản, dữ liệu được đưa vào giai đoạn **biến đổi thành dạng số học** để phục vụ cho quá trình huấn luyện mô hình học máy. Trong đồ án này, chúng tôi sử dụng kỹ thuật **tokenization** để thực hiện việc đó. Đây là quá trình tách văn bản thành các đơn vị nhỏ hơn gọi là token (thường là từ hoặc cụm từ), từ đó ánh xạ mỗi token thành một chỉ số số nguyên hoặc vector đặc trưng.

Sau khi hoàn thành các bước trên, dữ liệu đầu vào được chuẩn hóa ở mức cao, từ đó giúp các mô hình học sâu như Word2Vec, TF-IDF, LSTM hay Naive Bayes có thể khai thác hiệu quả thông tin ngữ nghĩa, nâng cao chất lượng phân tích cảm xúc.

## Huấn luyện mô hình

Trong bất kỳ hệ thống học máy nào, đặc biệt là các mô hình xử lý ngôn ngữ tự nhiên (NLP), bước tiền xử lý dữ liệu đầu vào đóng vai trò then chốt trong việc nâng cao hiệu quả và độ chính xác của mô hình. Dữ liệu thô thu thập được thường chứa nhiều thông tin dư thừa, không đồng nhất và có thể gây nhiễu nếu được sử dụng trực tiếp để huấn luyện. Do đó, việc làm sạch và chuẩn hóa dữ liệu là một bước không thể thiếu nhằm đảm bảo chất lượng tập huấn luyện.

Sơ đồ huấn luyện mô hình:



Hình 4.3 Sơ đồ huấn luyện

Quá trình xây dựng mô hình nhận diện cảm xúc trong văn bản tiếng Việt được triển khai theo một chuỗi các bước xử lý và huấn luyện mô hình học máy cũng như học sâu, như được minh họa trong sơ đồ trên (Hình 4.3).

**Bước 1: Tiền xử lý văn bản thô**

Bắt đầu từ một **tập văn bản tiếng Việt dung lượng lớn**, dữ liệu đầu vào được đưa qua công cụ **VNTokenizer** – một công cụ chuyên biệt cho việc **tách từ tiếng Việt**. Việc tách từ là bước đầu tiên trong chuỗi xử lý ngôn ngữ, giúp phân định rõ ràng ranh giới từ vựng, tạo tiền đề cho các bước biểu diễn và học máy phía sau.

**Bước 2: Biểu diễn văn bản**

Sau khi văn bản được tách từ, hệ thống áp dụng hai phương pháp biểu diễn đặc trưng văn bản để phục vụ cho hai nhánh xử lý song song:

* **Mô hình Word2Vec**: Được huấn luyện trên tập văn bản lớn, mô hình này ánh xạ mỗi từ sang một vector có ý nghĩa ngữ nghĩa trong không gian liên tục. Các vector từ sau đó được đưa vào mạng **LSTM** nhằm trích xuất các đặc trưng ngữ cảnh trong chuỗi văn bản.
* **Mô hình TF-IDF**: Một phương pháp biểu diễn văn bản cổ điển, TF-IDF chuyển đổi các văn bản thành các vector đặc trưng dựa trên tần suất từ, sau đó được sử dụng làm đầu vào cho các mô hình phân loại học máy truyền thống như Naive Bayes và SVM.

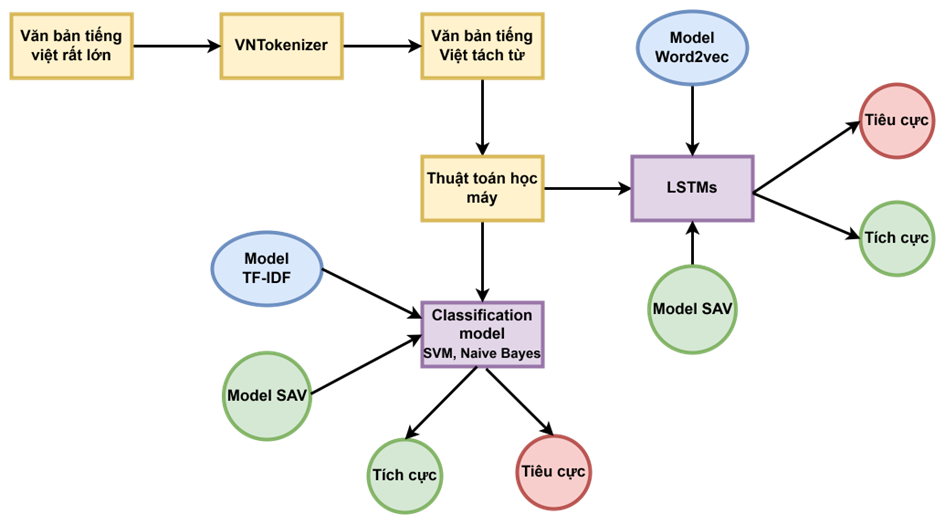
**Bước 3: Huấn luyện mô hình**

Với hai hướng biểu diễn dữ liệu đầu vào, hệ thống triển khai hai mô hình huấn luyện tương ứng:

* **Mạng LSTM (Long Short-Term Memory)**: Mô hình học sâu được sử dụng để học các đặc trưng tuần tự trong văn bản. LSTM nhận đầu vào là chuỗi vector từ Word2Vec và tiến hành huấn luyện để xây dựng mô hình phân lớp cảm xúc.
* **Mô hình phân loại học máy (Classification model)**: Sử dụng các thuật toán học có giám sát như SVM, Naive Bayes, các đặc trưng TF-IDF được học và tối ưu để tạo thành mô hình dự đoán cảm xúc.

**Bước 4: Xuất ra mô hình SAV**

Cả hai nhánh – học sâu và học máy – đều cho ra sản phẩm cuối là **mô hình nhận diện cảm xúc SAV**. Các mô hình này có thể được sử dụng để phân loại cảm xúc của các văn bản mới (tích cực/tiêu cực), phục vụ cho các ứng dụng như chatbot, phân tích phản hồi khách hàng hoặc hệ thống hỗ trợ ra quyết định.



Hình 4.4 Sơ đồ kiểm tra

Quá trình kiểm tra (test) mô hình nhận diện cảm xúc được thực hiện thông qua một chuỗi xử lý đầu vào văn bản, áp dụng các mô hình đã huấn luyện để đưa ra dự đoán cảm xúc, như minh họa trong Hình 4.3.

**Bước 1: Văn bản cần phân tích**

Giai đoạn test bắt đầu từ đầu vào là các **văn bản tiếng Việt trong tập test**.

**Bước 2: Tiền xử lý văn bản bằng VNTokenizer**

Tất cả văn bản đầu vào đều được **xử lý bằng VNTokenizer** – công cụ tách từ tiếng Việt, giúp chuẩn hóa văn bản thành dạng dễ xử lý hơn cho các mô hình phía sau.

**Bước 3: Biểu diễn văn bản đầu vào**

Từ văn bản đã tách từ, hệ thống triển khai **hai nhánh biểu diễn song song** tương ứng với các mô hình đã huấn luyện:

* **Nhánh 1 – TF-IDF**:  
  Văn bản được chuyển thành vector đặc trưng thông qua mô hình TF-IDF đã được huấn luyện từ trước. Vector này làm đầu vào cho các mô hình học máy (SVM hoặc Naive Bayes).
* **Nhánh 2 – Word2Vec**:  
  Văn bản được ánh xạ sang chuỗi vector ngữ nghĩa sử dụng mô hình Word2Vec. Chuỗi này tiếp tục được đưa vào mô hình LSTM đã huấn luyện.

**Bước 4: Dự đoán cảm xúc bằng các mô hình SAV**

Sau khi biểu diễn, văn bản được đưa vào các mô hình phân loại đã huấn luyện để dự đoán nhãn cảm xúc:

* **Mô hình LSTM SAV**  
  Dựa trên vector Word2Vec, LSTM đưa ra dự đoán văn bản là **tích cực** hay **tiêu cực**.
* **Mô hình SVM / NB SAV**  
  Dựa trên vector TF-IDF, các mô hình học máy SVM hoặc Naive Bayes cũng đưa ra phân loại cảm xúc tương ứng (**tích cực** hoặc **tiêu cực**).

**Bước 5: So sánh và đánh giá kết quả**

Để kiểm tra độ chính xác của mô hình, dự đoán được so sánh với nhãn thực tế (nếu có). Các chỉ số như **precision, recall, F1-score** được tính toán và in ra thông qua hàm classification\_report trong thư viện sklearn.

Kết quả cũng có thể được **so sánh giữa các mô hình** (LSTM, SVM, NB) để chọn mô hình có hiệu quả tốt nhất cho ứng dụng thực tế.

## Thực nghiệm và đánh giá kết quả

Sau khi hoàn tất bước tiền xử lý dữ liệu, sẽ tiến hành thực nghiệm mô hình trên hai phương pháp biểu diễn văn bản: **Word2Vec** và **TF-IDF**. Mỗi phương pháp được kết hợp với các thuật toán phân loại khác nhau nhằm so sánh, đánh giá hiệu quả nhận diện cảm xúc trong văn bản tiếng Việt.

Ba thuật toán học máy – gồm Naive Bayes, SVM và LSTM – được sử dụng để huấn luyện mô hình trên tập dữ liệu đã chuẩn hóa. Trong từng lần thực nghiệm, các chỉ số như **độ chính xác phân loại**, **thời gian huấn luyện**, và **tốc độ dự đoán** đều được ghi nhận nhằm phục vụ việc đánh giá khách quan và đề xuất cải tiến cho các giai đoạn nghiên cứu tiếp theo.

Việc huấn luyện và kiểm thử được triển khai trên môi trường lập trình Python với các thư viện học máy phổ biến như Scikit-learn, TensorFlow và Gensim. Toàn bộ thí nghiệm được thực hiện trên thiết bị có thông số kỹ thuật như sau:

* **Thiết bị sử dụng**: Acer Nitro 5 AN515-45
* **Bộ xử lý**: AMD Ryzen 5 5600H
* **Ổ cứng lưu trữ**: SSD 512 GB
* **Bộ nhớ RAM**: DDR4 8 GB (2 khe).
* **Ngôn ngữ lập trình**: Python 3.12.4

Với cấu hình như trên, thời gian huấn luyện mô hình có thể không tối ưu như trên các hệ thống GPU hiện đại, nhưng vẫn đảm bảo đủ khả năng đánh giá và kiểm thử hiệu năng của mô hình trên quy mô dữ liệu vừa phải.

Sau giai đoạn tiền xử lý và vector hóa dữ liệu bằng mô hình Word2Vec, chúng tôi xây dựng mô hình học sâu dựa trên kiến trúc mạng nơ-ron hồi tiếp (LSTM) với các lớp chính như sau:

Bảng 4.3 Danh sách các mô hình vector hóa và thuật toán phân loại được sử dụng trong thực nghiệm

|  |  |  |
| --- | --- | --- |
| **Thứ tự** | **Mô hình vector hóa** | **Phương pháp thực hiện** |
| **1** | BoW | LTSMs |
| **2** | TF-IDF | Linear SVM |
| **3** | TF-IDF | Naïve Bayes |

Bảng 4.4 Các lớp chính trong mô hình LTSMs

|  |  |  |
| --- | --- | --- |
| **Lớp (type)** | **Output Shape** | **Số tham số (Param #)** |
| **Embedding** | (?, 300, 300) | 6,597,900 (không huấn luyện) |
| **Masking** | (?, 300, 300) | 0 |
| **LSTM (200)** | (?, 200) | Tự động xây dựng |
| **Dense (2)** | (?, 2) | Tự động xây dựng |

* **Tổng số tham số:** 6,597,900
* **Tham số huấn luyện:** 0 (Embedding được cố định)

Huấn luyện mô hình:

* Tách tập train/validation: Tỷ lệ validation 20%
* Kích thước batch (số lượng mẫu đưa vào một lúc): 50
* Số vòng lặp (epoch): 5

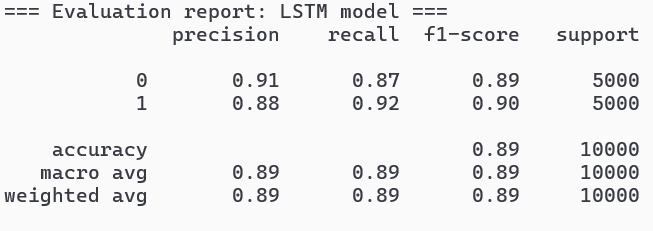
Kết quả sau mỗi epoch:

Bảng 4.5 Kết quả mỗi epoch huấn luyện

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Epoch | Accuracy (train) | Loss (train) | Accuracy (val) | Loss (val) |
| 1 | 0.7774 | 0.4648 | 0.8777 | 0.3003 |
| 2 | 0.8914 | 0.2769 | 0.8893 | 0.2741 |
| 3 | 0.9178 | 0.2119 | 0.8890 | 0.2798 |
| 4 | 0.9371 | 0.1640 | 0.8905 | 0.2951 |
| 5 | 0.9617 | 0.1098 | 0.8853 | 0.3135 |

**Phân loại bằng LSTM (trên tập dữ liệu kiểm tra)**

Báo cáo dựa trên tập dữ liệu kiểm tra về các chỉ số phân loạin chính:



Hình 4.5 Báo cáo trên tập dữ liệu kiểm tra với PP LSTM

**Phân tích chỉ số**

**1. Precision (Độ chính xác)**

* Lớp **0** (tiêu cực): 0.91 → khá cao, nghĩa là mô hình ít nhầm lẫn lớp tích cực thành tiêu cực.
* Lớp **1** (tích cực): 0.88 → cũng tốt, chỉ có một ít nhầm lẫn với lớp 0.
* Precision trung bình (macro/weighted): 0.89 → mô hình phân biệt các lớp khá ổn định.

**2. Recall (Khả năng phát hiện đúng)**

* Lớp **0**: 0.87 → mô hình vẫn còn bỏ sót một số văn bản tiêu cực.
* Lớp **1**: 0.92 → rất tốt, mô hình nhận diện văn bản tích cực khá đầy đủ.
* Trung bình: 0.89 → phản ánh độ phủ đều giữa 2 lớp.

**3. F1-score (Cân bằng giữa precision và recall)**

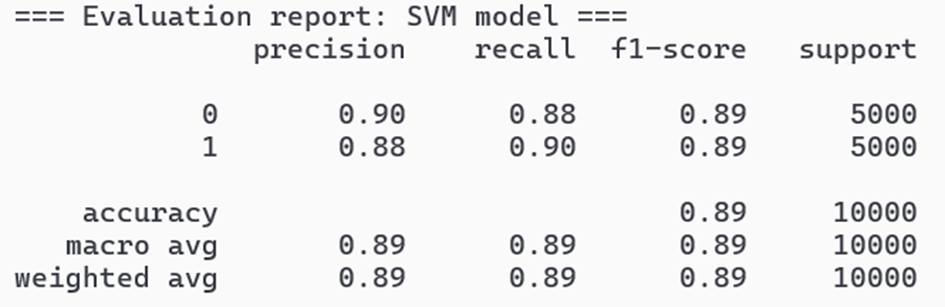
* F1-score cả hai lớp đều ~0.89–0.90 → cho thấy mô hình cân bằng tốt giữa việc phát hiện đúng và tránh nhầm lẫn.
* Chỉ số F1-score cao là dấu hiệu của mô hình có **độ chính xác tổng thể tốt**.

**4. Accuracy (Độ chính xác tổng thể)**

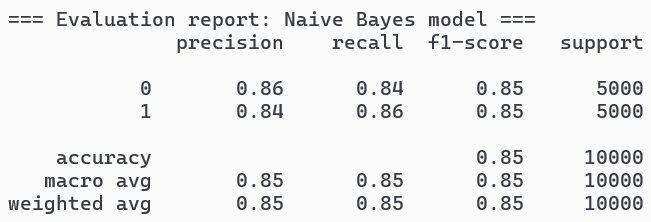
* Accuracy: **0.89** trên 10.000 mẫu kiểm tra.
* Đây là kết quả **khá ấn tượng**, đặc biệt nếu dữ liệu đầu vào đã được chuẩn hóa kỹ lưỡng và chia đều 2 lớp.

**Phân loại bằng mô hình học máy SVM và Naïve Bayes (trên tập dữ liệu kiểm tra):**

Các chỉ số như **precision, recall, F1-score** được tính toán và in ra thông qua hàm classification\_report trong thư viện sklearn.



Hình 4.6 Báo cáo trên tập dữ liệu kiểm tra với PP SVM



Hình 4.7 Báo cáo trên tập dữ liệu kiểm tra với PP Naïve Bayes

Dưới đây là phần **so sánh sơ bộ** giữa hai mô hình phân loại cảm xúc: **SVM** và **Naive Bayes**, dựa trên các chỉ số đánh giá cơ bản như precision, recall, f1-score và accuracy.

Về **precision** – tức độ chính xác của mô hình trong việc dự đoán đúng các nhãn:

* Mô hình SVM đạt precision là 0.90 cho nhãn "0" (tiêu cực) và 0.88 cho nhãn "1" (tích cực).
* Trong khi đó, mô hình Naive Bayes đạt lần lượt là 0.86 và 0.84 cho hai nhãn tương ứng.

Điều này cho thấy SVM có khả năng đưa ra các dự đoán chính xác hơn về cảm xúc so với Naive Bayes.

Xét đến **recall** – tức khả năng thu hồi, hay tỷ lệ các văn bản thật sự thuộc một nhãn được mô hình nhận diện đúng:

* SVM đạt recall là 0.88 với nhãn tiêu cực và 0.90 với nhãn tích cực.
* Naive Bayes thấp hơn một chút, với 0.84 và 0.86.

SVM vẫn chiếm ưu thế, cho thấy mô hình này ít bỏ sót các mẫu đúng hơn so với NB.

Về chỉ số **f1-score**, vốn là thước đo cân bằng giữa precision và recall:

* SVM giữ mức f1-score là 0.89 cho cả hai nhãn.
* Naive Bayes chỉ đạt 0.85 cho mỗi nhãn.

Điều này xác nhận rằng SVM cân bằng tốt hơn giữa việc dự đoán đúng và không bỏ sót các mẫu.

Về tổng thể, **độ chính xác toàn bộ (accuracy)** của SVM là 0.89, trong khi Naive Bayes là 0.85. Đây là sự khác biệt tương đối rõ ràng cho thấy SVM làm việc hiệu quả hơn trên tập dữ liệu kiểm tra. Ngoài ra, trung bình cộng các chỉ số (macro avg) và trung bình có trọng số (weighted avg) của SVM cũng cao hơn hẳn so với Naive Bayes.

KẾT LUẬN

Sau quá trình khảo sát, nghiên cứu, cài đặt và thử nghiệm, dưới sự hướng dẫn tận tình của thầy Kiều Tuấn Dũng, em đã hoàn thành hệ thống quản lý phòng thực hành máy tính cho Trường Đại học Thủy lợi. Hệ thống này được xây dựng nhằm đáp ứng nhu cầu quản lý và tối ưu hóa việc sử dụng các phòng thực hành máy tính, bao gồm các chức năng như điểm danh sinh viên, báo cáo sự cố thiết bị, quản lý tình trạng thiết bị và cung cấp các thống kê, báo cáo tổng quan của phòng thực hành máy tính.

Về mặt công nghệ, em đã nắm chắc và áp dụng thành công các công cụ và phần mềm vào việc phát triển website. Hệ thống được xây dựng với PHP Laravel cho backend, đảm bảo ổn định và hiệu quả trong xử lý dữ liệu. Giao diện người dùng sử dụng HTML, CSS, JavaScript, kết hợp với Bootstrap, jQuery và Ajax, giúp hệ thống thân thiện và dễ sử dụng. Hệ thống đã trải qua kiểm thử với các kịch bản khác nhau để đảm bảo tính logic và ổn định khi vận hành.

Hệ thống đã được xây dựng hoàn chỉnh và đáp ứng đầy đủ các chức năng theo mục tiêu của bản đề cương đồ án tốt nghiệp. Do hệ thống tập trung vào việc quản lý phòng máy, nên các chức năng quản lý tài khoản sinh viên, tài khoản giảng viên, lớp học phần và cơ sở vật chất đã được chuyển giao cho các kỹ thuật viên đảm nhiệm. Mặc dù hệ thống đã đạt được các mục tiêu đề ra, vẫn còn nhiều khía cạnh cần tiếp tục nghiên cứu và hoàn thiện để nâng cao hiệu quả và chất lượng của hệ thống.

**Hướng phát triển đề tài:**

* Nghiên cứu và chi tiết hóa thống kê báo cáo tình trạng sử dụng phòng máy.
* Phát triển thêm tính năng chia sẻ màn hình của sinh viên trong tiết học.

Kết quả của đề tài này không chỉ mang lại cái nhìn tổng quan về quản lý phòng thực hành máy tính mà còn trang bị cho em những kiến thức và kỹ năng quan trọng cho công việc sau này. Em rất mong nhận được những góp ý, chỉ dẫn từ thầy cô để nâng cao hơn nữa chất lượng và hiệu quả của hệ thống.

Em xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | H. Gomaa, *Software Modeling and Design: UML, Use Cases, Patterns, and Software Architectures*, Cambridge University Press, 2011. |
| [2] | Trương Ninh Thuận, Đặng Đức Hạnh, *Giáo trình phân tích và Thiết kế hướng đối tượng*, Đại học Quốc gia Hà Nội, 2013. |
| [3] | M. Said, *Laravel Queues in Action*, Self, 2020. |
| [4] | I. Sommerville, *Software Engineering*, 10th ed., Pearson, 2015. |
| [5] | M. L. Gillenson, *Fundamentals of Database Management Systems*, 2nd ed., John Wiley & Sons Inc, 2011. |
| [6] | J. Duckett, *HTML and CSS: Design and Build Websites*, John Wiley & Sons, 2011. |
| [7] | J. Duckett, *JavaScript and jQuery: Interactive Front-End Web Development*, Wiley, 2014. |
| [8] | Martin Fowler, Kent Beck, John Brant, William Opdyke, Don Roberts, *Refactoring: Improving the Design of Existing Code*, 2nd ed., Addison-Wesley Professional, 2018. |
| [9] | J. Duckett, *PHP & MySQL: Server-side Web Development*, Wiley, 2022. |
| [10] | M. Stauffer, *Laravel Up & Running: A Framework for Building Modern PHP Apps*, 3rd ed., O'Reilly Media, 2023. |

PHỤ LỤC