LỜI MỞ ĐẦU

LỜI CẢM ƠN

NHẬN XÉT
(Của giảng viên hướng dẫn)

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Giảng viên hướng dẫn
(Ký và ghi rõ họ tên)

UBND TỈNH TRÀ VINH
TRƯỜNG ĐẠI HỌC TRÀ VINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(Của giảng viên hướng dẫn)

Họ và tên sinh viên:…… MSSV: ……

Ngành:…… Khóa:……..

Tên đề tài:…….

……..

……..

Họ và tên Giảng viên hướng dẫn:…….

Chức danh:……... Học vị:……

NHẬN XÉT

1. Nội dung đề tài:

……

……

……

……

……

……

……

2. Ưu điểm:

……

……

……

3. Khuyết điểm:

……

……

……

……

4. Điểm mới đề tài:

……
……
……
……

5. Giá trị thực trên đề tài:

……
……
……
……

6. Đề nghị sữa chửa bổ sung:

……
……
……
……
……

7. Đánh giá

……
……
……
……

Trà Vinh, *ngày tháng năm 2025*
Giảng viên hướng dẫn
(Ký và ghi rõ họ tên)

NHẬN XÉT
(Của giảng viên chấm)

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Giảng viên chấm
(Ký và ghi rõ họ tên)

UBND TỈNH TRÀ VINH
TRƯỜNG ĐẠI HỌC TRÀ VINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(Của giảng viên chấm)

Họ và tên người nhận xét:……

Chức danh:……... Học vị:……

Chuyên ngành:……..

Cơ quan công tác:…….

Tên sinh viên:……

Tên đề tài:…….

……..

……..

I. Ý KIẾN NHẬN XÉT

1. Nội dung:

……

……

……

……

……

……

……

……

2. Điểm mới các kết quả của khóa luận:

……

……

……

3. Ứng dụng thực tế:

……

……

……

……

II. CÁC VẤN ĐỀ CẦN LÀM RÕ

(Các câu hỏi của giáo viên phản biện)

……

……

……

……

……

……

……

……

III. KẾT LUẬN

(Ghi rõ đồng ý hay không đồng ý cho bảo vệ đồ án khóa luận tốt nghiệp)

……

……

……

……

……

Trà Vinh, *ngày tháng năm 2025*

Giảng viên chấm

(Ký và ghi rõ họ tên)

MỤC LỤC

DANH MỤC CÁC BẢNG, SƠ ĐỒ, HÌNH

DANH MỤC CÁC CỤM TỪ VIẾT TẮT

CHƯƠNG 1. TỔNG QUAN

1.1 Tổng quan đề tài

Trong thời đại số hóa hiện tại, đặc biệt là với sự phát triển mạnh mẽ của các nền tảng thương mại điện tử, người tiêu dùng đang có xu hướng chuyển từ hình thức mua sắm trực tiếp sang mua sắm trực tuyến về đa số hàng hóa tiêu dùng. Vì vậy, các cửa hàng, doanh nghiệp cũng đang mở rộng kinh doanh trực tuyến để đáp ứng được nhu cầu của khách hàng.

Qua đó, khi các cửa hàng, doanh nghiệp có một lượng lớn khách hàng và doanh số sản phẩm bán ra được, việc quản lý, chăm sóc lấy ý kiến đánh giá từ khách hàng là rất quan trọng. Nhưng vì khi có lượng lớn ý kiến đánh giá như vậy, việc phân tích theo dõi và tiếp nhận những ý kiến đó theo cách thủ công truyền thống là vô cùng khó khăn. Vì thế, cần một giải pháp thực tiễn nào đó để việc tiếp nhận được ý kiến của lượng lớn khách hàng một cách nhanh chống và hoàn toàn tự động.

Cùng với đó, việc phát triển của trí tuệ nhân tạo những năm gần đây mở ra một ý tưởng trong việc tự động hóa các tác vụ vốn trước đây cần đến sự can thiệp thủ công, trong đó có bài toán phân tích và phân loại cảm xúc từ phản hồi của khách hàng. Với sự hỗ trợ của các kỹ thuật xử lý ngôn ngữ tự nhiên kết hợp cùng các mô hình học máy và học sâu, hoàn toàn có thể xây dựng một hệ thống tự động có khả năng hiểu và phân loại cảm xúc lượng lớn ý kiến đánh giá của khách hàng một cách nhanh chóng, chính xác và hiệu quả.

Đặc biệt, trong bối cảnh ngôn ngữ tiếng Việt vẫn còn hạn chế về tài nguyên xử lý ngôn ngữ so với các ngôn ngữ lớn như tiếng Anh, việc phát triển các hệ thống phân tích cảm xúc dành riêng cho tiếng Việt mang ý nghĩa rất quan trọng cả về mặt ứng dụng lẫn nghiên cứu. Không chỉ giúp cửa hàng, doanh nghiệp nắm bắt nhanh ý kiến cảm nhận của khách hàng, cải thiện chất lượng sản phẩm, dịch vụ mà còn góp phần thúc đẩy sự phát triển của các công nghệ ngôn ngữ tự nhiên cho tiếng Việt.

Vì vậy, "Phân loại cảm xúc tiếng Việt bằng mô hình NLP" không chỉ có giá trị trong phạm vi học thuật mà còn mang tính thực tiễn cao. Nó hướng đến việc

ứng dụng công nghệ hiện đại nhằm giải quyết bài toán thực tế đang được quan tâm, đồng thời là cơ hội để người học tiếp cận và làm chủ các công cụ, kỹ thuật mới trong lĩnh vực trí tuệ nhân tạo, lĩnh vực đang ngày càng phát triển mạnh mẽ và đầy tiềm năng trong thời đại số hóa hiện nay.

1.2 Mục tiêu đề tài

Mục tiêu chính của đề tài "Phân loại cảm xúc tiếng Việt bằng mô hình NLP" là nghiên cứu và xây dựng một hệ thống có khả năng tự động phân tích và phân loại cảm xúc từ các bình luận tiếng Việt, cụ thể là các bình luận người dùng trên các nền tảng thương mại điện tử. Hệ thống này cần có khả năng nhận diện và phân loại chính xác bình luận thành ba nhóm cảm xúc: tích cực, tiêu cực và trung lập, từ đó hỗ trợ các cửa hàng và doanh nghiệp hiểu rõ hơn về cảm nhận của khách hàng đối với sản phẩm, dịch vụ của mình.

Về mục tiêu học tập, nghiên cứu:

Nhằm củng cố và mở rộng kiến thức, kỹ năng chuyên môn trong lĩnh vực trí tuệ nhân tạo, đặc biệt là xử lý ngôn ngữ tự nhiên và học máy.

Hiểu và áp dụng kiến thức về xử lý ngôn ngữ tự nhiên tiếng Việt: Tìm hiểu sâu về đặc trưng ngôn ngữ tiếng Việt, bao gồm tách từ, chuẩn hóa văn bản, xử lý văn bản, đặc biệt là làm việc với ngôn ngữ thực tế trong các bình luận.

Làm quen và sử dụng các mô hình học máy và học sâu trong xử lý ngôn ngữ tự nhiên: Học cách sử dụng và so sánh hiệu quả giữa các thuật toán học máy truyền thống như Naive Bayes và các mô hình học sâu tiên tiến như BERT, đặc biệt là các mô hình tiền huấn luyện dành riêng cho tiếng Việt như PhoBERT.

Cải thiện khả năng nghiên cứu khoa học và tư duy phản biện: Tìm hiểu, phân tích tài liệu nghiên cứu, đánh giá kết quả thực nghiệm, so sánh mô hình, và đưa ra giải pháp cải tiến là những bước quan trọng giúp rèn luyện tư duy khoa học và nâng cao năng lực nghiên cứu độc lập.

1.3 Phạm vi nghiên cứu

Đề tài tập trung vào việc ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên, học máy và học sâu để tự động phân loại cảm xúc trong các bình luận tiếng Việt.

Cụ thể, phạm vi của đề tài được giới hạn trong ngữ cảnh phân tích cảm xúc của người dùng khi phản hồi sản phẩm trên các nền tảng thương mại điện tử phổ biến tại Việt Nam. Đây là môi trường dữ liệu thực tế, nơi người dùng thường xuyên để lại nhận xét, đánh giá về chất lượng sản phẩm, dịch vụ đã sử dụng. Việc lựa chọn dữ liệu từ các trang thương mại điện tử giúp đề tài tiếp cận sát với nhu cầu ứng dụng trong thực tiễn, đồng thời mang đến những thách thức thú vị trong quá trình xử lý ngôn ngữ tự nhiên tiếng Việt.

Dữ liệu của đề tài là các bình luận dạng văn bản, được trích xuất từ trang thương mại điện tử Tiki. Cảm xúc của các bình luận sẽ được phân loại thành ba nhóm cơ bản: tích cực, tiêu cực và trung lập. Đây là cách phân loại cảm xúc đơn giản nhưng hiệu quả, phù hợp với bài toán phân tích cảm xúc ở cấp độ ứng dụng. Việc gán nhãn được thực hiện một cách thủ công, tổng hợp từ nhiều người.

Đề tài áp dụng các bước tiền xử lý văn bản phù hợp với tiếng Việt, bao gồm: chuẩn hóa chữ viết, tách từ, loại bỏ ký tự đặc biệt, dấu câu. Sau khi tiền xử lý, dữ liệu sẽ được đưa vào các mô hình học máy và học sâu để huấn luyện. Các mô hình truyền thống như Naive Bayes được sử dụng như cơ sở so sánh, sau đó mô hình PhoBERT sẽ được áp dụng để nâng cao độ chính xác. Việc huấn luyện và đánh giá mô hình được thực hiện trên máy tính cá nhân cùng với nền tảng Google Colab.

1.4 Phương pháp nghiên cứu

Để đạt được mục tiêu của đề tài, quá trình thực hiện sẽ được chia thành các bước cụ thể, tuần tự theo quy trình xử lý dữ liệu và phát triển mô hình học máy và học sâu trong lĩnh vực xử lý ngôn ngữ tự nhiên. Cụ thể, các phương pháp được áp dụng bao gồm:

Khảo sát lý thuyết và tổng quan nghiên cứu: Tìm hiểu các khái niệm liên quan đến phân tích cảm xúc, các cấp độ cảm xúc và các loại cảm xúc phổ biến. Nghiên cứu các phương pháp dựa trên học máy và học sâu. Khảo sát các công trình nghiên cứu trước đó liên quan đến phân tích cảm xúc trong tiếng Việt và các ngôn ngữ khác.

Thu thập và xử lý dữ liệu

Thu thập dữ liệu các bình luận của người dùng trên các nền tảng thương mại điện tử phổ biến tại Việt Nam như Shopee, Tiki và Lazada. Các bình luận này được viết bằng tiếng Việt, với văn tự nhiên, đôi khi có cả các từ viết tắt hoặc không có dấu câu. Điều này mang lại tính thực tế cao, phản ánh đúng với thực tế trong việc ứng dụng bài toán phân loại cảm xúc.

Sau khi thu thập dữ liệu, tiếp theo là xử lý và gán nhãn cảm xúc cho từng bình luận. Xử lý loại bỏ các ký tự đặc biệt, các biểu tượng cảm xúc, các dấu câu,… Để phục vụ cho mục tiêu nghiên cứu, mỗi bình luận sẽ được phân loại vào một trong ba nhãn cảm xúc: tích cực, tiêu cực, hoặc trung lập theo các tiêu chí riêng về các từ cảm xúc và được thực hiện bởi nhiều người, sau đó tổng hợp thống nhất về nhãn của dữ liệu.

Xây dựng mô hình học máy và học sâu: Thử nghiệm các mô hình học máy cổ điển để làm cơ sở so sánh, triển khai các mô hình học sâu phổ biến cho xử lý chuỗi văn bản, tinh chỉnh tham số và tối ưu mô hình.

Huấn luyện và đánh giá mô hình: Chia dữ liệu thành các tập huấn luyện, kiểm thử và kiểm tra theo tỷ lệ hợp lý, sử dụng các thước đo đánh giá như Accuracy, Precision, Recall, F1-score, Confusion Matrix để đo lường hiệu quả mô hình, so sánh kết quả giữa các mô hình học máy, học sâu và mô hình tiền huấn luyện để chọn ra phương án tối ưu.

Triển khai ứng dụng thử nghiệm: Phát triển một ứng dụng web đơn giản cho phép người dùng nhập vào đường dẫn của sản phẩm trên các sàn thương mại điện tử như Shopee, Tiki, Lazada, sau đó hệ thống sẽ tự động đánh giá các bình luận của sản phẩm đó và trả kết quả sau khi phân tích về cho người dùng, cụ thể ở đây là ứng dụng cho các cửa hàng và doanh nghiệp.

Tổng hợp kết quả và đề xuất hướng phát triển: Đưa ra nhận xét, phân tích những ưu, nhược điểm của các mô hình đã thử nghiệm. So sánh hiệu quả mô hình với các kết quả trong các công trình nghiên cứu khác. Đề xuất hướng phát

triển tiếp theo như mở rộng phạm vi cảm xúc, áp dụng cho các lĩnh vực khác như y tế, giáo dục,...

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Tổng quan về xử lý ngôn ngữ tự nhiên

2.1.1 Khái niệm

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực liên ngành giữa khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học, nhằm mục tiêu xây dựng các hệ thống có khả năng hiểu, diễn giải và tạo ra ngôn ngữ của con người một cách tự động. NLP đóng vai trò trung gian trong việc giao tiếp giữa con người và máy tính bằng ngôn ngữ tự nhiên, cho phép máy tính xử lý và phân tích lượng lớn văn bản, lời nói mà không cần đến các lệnh lập trình phức tạp từ phía người dùng.

Với sự phát triển mạnh mẽ của dữ liệu văn bản trên các nền tảng số như mạng xã hội, thương mại điện tử, diễn đàn trực tuyến,... NLP ngày càng trở thành một công cụ không thể thiếu trong việc phân tích dữ liệu phi cấu trúc. Các ứng dụng phổ biến của NLP có thể kể đến như: phân tích cảm xúc, chatbot, dịch máy, trích xuất thông tin, tóm tắt văn bản, nhận dạng thực thể, và nhiều ứng dụng khác.

Trong những năm gần đây, sự phát triển của các mô hình ngôn ngữ lớn như BERT, GPT,... đã mang lại bước tiến vượt bậc cho NLP, đặc biệt là trong các bài toán phức tạp như phân loại cảm xúc, dịch máy, và hiểu ngữ cảnh sâu. Đối với tiếng Việt, mô hình PhoBERT được huấn luyện riêng cho dữ liệu tiếng Việt đã chứng minh hiệu quả vượt trội trong nhiều tác vụ, bao gồm phân tích cảm xúc.

2.1.2 Quy trình xử lý ngôn ngữ tự nhiên

Quy trình xử lý ngôn ngữ tự nhiên thường bao gồm nhiều giai đoạn:

Thu thập và tiền xử lý dữ liệu văn bản: Dữ liệu văn bản ban đầu thường chứa nhiều yếu tố không cần thiết như dấu câu, ký tự đặc biệt, từ dừng, từ viết tắt, lỗi chính tả,... nên cần được làm sạch và chuẩn hóa. Đối với tiếng Việt, bước tách từ là cực kỳ quan trọng do tiếng Việt là ngôn ngữ đơn âm tiết, không có dấu cách rõ ràng giữa các từ như tiếng Anh.

Biểu diễn văn bản: Sau khi tiền xử lý, văn bản cần được chuyển đổi thành dạng số để mô hình có thể xử lý được. Một số kỹ thuật phổ biến gồm có: Bag of Words (BoW), TF-IDF, word embeddings (Word2Vec, GloVe)…

Xây dựng mô hình học máy và học sâu: Tùy theo bài toán và yêu cầu về độ chính xác, người dùng có thể lựa chọn mô hình học máy truyền thống như Naive Bayes, SVM, Decision Tree,... hoặc các mô hình học sâu như CNN, RNN, LSTM, Transformer.

Đánh giá và tối ưu mô hình: Các mô hình sau khi huấn luyện cần được đánh giá bằng các chỉ số như độ chính xác (accuracy), độ bao phủ (recall), độ chính xác (precision), F1-score,... để so sánh và lựa chọn mô hình tốt nhất.

2.2 Phân tích cảm xúc

2.2.1 Khái niệm

Phân tích cảm xúc (Sentiment Analysis) hay còn gọi là phân loại cảm xúc, là một nhánh quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định thái độ, quan điểm hoặc cảm xúc của người viết đối với một chủ đề cụ thể trong văn bản. Mục tiêu chính của phân tích cảm xúc là xác định liệu văn bản thể hiện cảm xúc tích cực, tiêu cực hay trung lập, từ đó giúp các hệ thống hiểu được "ý định" hoặc "cảm nhận" của người dùng khi giao tiếp bằng ngôn ngữ tự nhiên.

Phân tích cảm xúc có rất nhiều ứng dụng trong thực tế, đặc biệt là trong lĩnh vực kinh doanh và chăm sóc khách hàng. Các doanh nghiệp có thể sử dụng phân tích cảm xúc để theo dõi phản hồi của người tiêu dùng, đánh giá mức độ hài lòng về sản phẩm, dịch vụ hoặc thương hiệu, từ đó đưa ra chiến lược cải thiện phù hợp. Ngoài ra, phân tích cảm xúc còn được ứng dụng trong phân tích dư luận xã hội, đánh giá hiệu quả chiến dịch truyền thông, quản lý danh tiếng và nhiều lĩnh vực khác.

Về mặt kỹ thuật, phân tích cảm xúc có thể được triển khai dưới nhiều mức độ khác nhau:

Mức câu: Xác định cảm xúc tổng thể của từng câu.

Mức đoạn văn, bình luận: Đánh giá cảm xúc chung của toàn bộ văn bản hoặc bài viết.

Mức thực thể: Phân tích cảm xúc hướng đến các khía cạnh cụ thể (ví dụ: người dùng thích "giá cả" nhưng không hài lòng với "chất lượng sản phẩm").

Đối với ngôn ngữ tiếng Việt, phân tích cảm xúc là một thách thức lớn do đặc điểm ngữ pháp đặc biệt, cách viết không chuẩn, sử dụng từ láy, tiếng lóng, viết tắt, cũng như hiện tượng từ đồng âm, dị nghĩa phổ biến. Do đó, việc áp dụng các mô hình hiện đại được thiết kế riêng cho tiếng Việt như PhoBERT sẽ mang lại kết quả tốt hơn so với việc sử dụng các mô hình học từ tiếng Anh hoặc ngôn ngữ khác.

Phân tích cảm xúc là một công cụ hữu ích giúp biến đổi dữ liệu văn bản phi cấu trúc thành thông tin có giá trị. Trong thời đại số hóa và dữ liệu lớn, việc áp dụng các kỹ thuật phân tích cảm xúc không chỉ mang lại lợi thế cạnh tranh cho doanh nghiệp mà còn mở ra nhiều hướng nghiên cứu và ứng dụng trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên.

2.2.2 Các phương pháp phân tích cảm xúc

Phương pháp dựa trên từ điển cảm xúc: Sử dụng một tập hợp từ điển chứa các từ có gán nhãn cảm xúc (ví dụ: “tuyệt vời” là tích cực, “tệ” là tiêu cực). Văn bản được phân tích dựa trên việc đếm và tổng hợp các từ có trong từ điển để đưa ra dự đoán về cảm xúc. Tuy đơn giản và dễ triển khai, phương pháp này thường thiếu hiệu quả trong việc xử lý các ngữ cảnh phức tạp hoặc câu mang tính mỉa mai, ẩn dụ.

Phương pháp học máy truyền thống: Bao gồm các mô hình như Naive Bayes, SVM, KNN,... Các mô hình này học từ dữ liệu huấn luyện đã được gán nhãn để rút ra quy tắc phân loại. Trước khi đưa vào mô hình, văn bản thường được chuyển thành dạng vector thông qua các kỹ thuật biểu diễn như TF-IDF hoặc Bag of Words.

Phương pháp học sâu: Các mô hình học sâu như CNN, RNN, LSTM hoặc Transformer (BERT, RoBERTa, PhoBERT) cho phép khai thác tốt hơn mối quan hệ ngữ cảnh và cấu trúc ngôn ngữ trong văn bản. Đặc biệt, các mô hình tiền huấn

luyện như BERT hoặc PhoBERT có khả năng hiểu ngữ nghĩa theo ngữ cảnh sâu, giúp cải thiện đáng kể độ chính xác trong phân tích cảm xúc, đặc biệt đối với ngôn ngữ nhiều biến thể như tiếng Việt.

2.3 Kỹ thuật biểu diễn văn bản

Trong xử lý ngôn ngữ tự nhiên, việc biểu diễn văn bản dưới dạng mà máy tính có thể hiểu và xử lý được là một bước quan trọng, đặc biệt trong các bài toán phân loại văn bản như phân loại cảm xúc. Văn bản ở dạng tự nhiên là một chuỗi ký tự hoặc từ ngữ, nhưng để xử lý bằng các thuật toán học máy hoặc học sâu, văn bản cần được chuyển đổi sang dạng số hoặc vector đặc trưng.

2.3.1 Kỹ thuật Bag of Word (BoW)

Bag of Words là một trong những kỹ thuật biểu diễn văn bản đơn giản và phổ biến nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên. Ý tưởng chính của BoW là đại diện cho một văn bản như một tập hợp các từ ngữ (không quan tâm đến thứ tự) và biểu diễn văn bản dưới dạng vector dựa trên tần suất xuất hiện của từng từ trong tập từ vựng.

Cụ thể, với một tập hợp các văn bản, trước tiên người ta xây dựng một tập từ vựng bao gồm tất cả các từ xuất hiện trong tập dữ liệu huấn luyện. Mỗi văn bản sau đó sẽ được biểu diễn bằng một vector có chiều bằng kích thước của tập từ vựng. Mỗi phần tử trong vector đại diện cho số lần xuất hiện của một từ cụ thể trong văn bản đó. Ví dụ, nếu từ "vui" xuất hiện ba lần trong văn bản, thì phần tử tương ứng với từ "vui" trong vector sẽ có giá trị là 3.

BoW có ưu điểm là dễ hiểu, dễ triển khai và hiệu quả trong các bài toán phân loại cơ bản. Tuy nhiên, kỹ thuật này tồn tại nhiều hạn chế. Trước hết, nó bỏ qua thứ tự từ trong câu, do đó không nắm bắt được ngữ nghĩa đầy đủ của văn bản. Ngoài ra, BoW không thể hiện được mối quan hệ ngữ cảnh giữa các từ, ví dụ như sự khác biệt giữa "không vui" và "rất vui" có thể bị mất đi khi chỉ xét đến tần suất của từ "vui".

Một vấn đề khác là độ hiệu quả tính toán: với các tập từ vựng lớn, vector BoW sẽ có kích thước rất cao và thường rất thưa, gây tốn bộ nhớ và ảnh hưởng đến tốc độ xử lý.

Mặc dù có nhiều điểm hạn chế, BoW vẫn là một kỹ thuật nền tảng quan trọng trong lĩnh vực NLP và thường được dùng như một bước khởi đầu để so sánh với các mô hình phức tạp hơn như TF-IDF hay Word Embedding. Trong các bài toán phân loại cảm xúc tiếng Việt, BoW có thể hoạt động tốt với dữ liệu nhỏ và khi kết hợp với các kỹ thuật tiền xử lý, chuẩn hóa từ hoặc phân tách từ chính xác.

2.3.2 Kỹ thuật TF-IDF

TF-IDF (Term Frequency - Inverse Document Frequency) là một kỹ thuật biểu diễn văn bản nhằm đánh giá mức độ quan trọng của một từ trong một văn bản so với toàn bộ tập hợp các văn bản. Đây là sự cải tiến của mô hình Bag of Words, với mục tiêu không chỉ xét đến tần suất xuất hiện của từ trong văn bản, mà còn điều chỉnh dựa trên mức độ phổ biến của từ đó trên toàn bộ tập dữ liệu.

Cụ thể, TF-IDF là tích của hai thành phần:

TF (Term Frequency): đo tần suất xuất hiện của từ trong văn bản. Có thể tính bằng công thức:

TF(t, d) = (số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong d)

IDF (Inverse Document Frequency): đo mức độ đặc trưng của từ, bằng cách giảm trọng số của các từ xuất hiện nhiều trong toàn bộ tập văn bản. Công thức thường dùng là:

$$IDF(t) = \log(N / (df(t) + 1))$$

Trong đó, N là tổng số văn bản trong tập dữ liệu, còn df(t) là số văn bản chứa từ t. Việc cộng thêm 1 vào mẫu số giúp tránh chia cho 0.

Khi nhân TF với IDF, ta được giá trị TF-IDF thể hiện độ quan trọng tương đối của từ trong một văn bản cụ thể. Các từ phổ biến như “và”, “là”, “của” có thể có TF cao nhưng do xuất hiện nhiều trong hầu hết các văn bản nên sẽ có IDF thấp, dẫn đến TF-IDF thấp. Ngược lại, các từ hiếm nhưng mang tính phân biệt cao (ví

dụ: “phấn khích”, “tức giận”, “buồn bã”) sẽ có trọng số cao, từ đó giúp mô hình học được những đặc điểm đặc trưng của từng cảm xúc.

Ưu điểm chính của TF-IDF là khả năng làm nổi bật các từ khóa quan trọng và loại bỏ các từ mang tính chung chung. Trong các bài toán phân loại cảm xúc, TF-IDF giúp mô hình nhận diện được các từ thể hiện cảm xúc rõ ràng, từ đó nâng cao hiệu quả phân loại.

Tuy nhiên, tương tự như Bag of Words, TF-IDF không xét đến ngữ cảnh hoặc thứ tự từ trong câu. Điều này có thể gây khó khăn trong việc hiểu các cấu trúc ngữ nghĩa phức tạp như phủ định ("không vui") hoặc các cụm từ có sắc thái cảm xúc đặc biệt. Ngoài ra, biểu diễn văn bản dưới dạng TF-IDF vẫn tạo ra vector thưa và kích thước cao, gây tốn tài nguyên tính toán với các tập dữ liệu lớn.

Mặc dù vậy, TF-IDF vẫn là một kỹ thuật mạnh mẽ và thường được dùng làm baseline trong các mô hình phân loại văn bản, bao gồm cả phân loại cảm xúc tiếng Việt. Khi kết hợp với các thuật toán học máy truyền thống như Naive Bayes, SVM hay Logistic Regression, TF-IDF có thể đạt hiệu quả tốt trên các tập dữ liệu vừa và nhỏ.

2.4 Tổng quan về học máy và học sâu

2.5 Các mô hình dùng trong xử lý ngôn ngữ tự nhiên