ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN
MÔN XỬ LÝ THÔNG TIN GIỌNG NÓI – DS313.O21
Đề tài: PHÁT HIỆN CẢM XÚC TỪ GIỌNG NÓI

GVHD: ThS. Nguyễn Thành Luân
Nhóm 5:

1. Trần Hoài Bão MSSV: 21250628

Tp. Hồ Chí Minh, 05/2024

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

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

....., ngày…….....tháng……năm 2024

Người nhận xét
(Ký tên và ghi rõ họ tên)

BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN:

Bảng 1: Bảng phân công, đánh giá thành viên

Họ và tên	MSSV	Phân công	Đánh giá
Trần Hoài Bão	21520628	Tuần 1: Tìm hiểu bài toán, tiền xử lý dữ liệu Tuần 2: Áp dụng mô hình học máy, kiểm tra kết quả Tuần 3: Viết báo cáo	Tuần 1: đạt Tuần 2: đạt Tuần 3: đạt
XXX	XXX	Tuần 1: Tuần 2: Tuần 3:	Tuần 1: Tuần 2: Tuần 3:
XXX	XXX	Tuần 1: Tuần 2: Tuần 3:	Tuần 1: Tuần 2: Tuần 3:
XXX	XXX	Tuần 1: Tuần 2: Tuần 3:	Tuần 1: Tuần 2: Tuần 3:

LỜI MỞ ĐẦU

DANH MỤC CÁC BẢNG, HÌNH ẢNH

Danh mục các bảng:

Danh mục hình ảnh:

Mục lục

Chương 1: GIỚI THIỆU CHUNG

1. **Giới thiệu môn học**

   Môn học "Xử lý thông tin giọng nói" là một trong những môn học chuyên ngành liên quan đến xử lý ngôn ngữ tự nhiên và công nghệ thông tin. Trong thời đại hiện đại, sự phát triển của các công nghệ về truyền thông âm thanh và xử lý ngôn ngữ đã mở ra những triển vọng vô cùng hứa hẹn cho việc ứng dụng của xử lý thông tin giọng nói trong nhiều lĩnh vực khác nhau.

   Môn học này tập trung vào nghiên cứu và ứng dụng các kỹ thuật xử lý tín hiệu âm thanh và phân tích ngôn ngữ nhằm giải quyết các vấn đề thực tiễn trong cuộc sống và các lĩnh vực khoa học kỹ thuật. Sinh viên sẽ được giới thiệu về các khái niệm cơ bản của xử lý thông tin giọng nói, từ đó tiếp cận và hiểu rõ các vấn đề phức tạp hơn như nhận dạng giọng nói, tổng hợp giọng nói, trích xuất đặc trưng âm thanh, và ứng dụng của chúng trong các lĩnh vực như hệ thống nhận dạng giọng nói, trợ lý ảo, và giao tiếp ngôn ngữ tự nhiên.

   Môn học "Xử lý thông tin giọng nói" còn đặt ra các thách thức đáng giá và cung cấp cho sinh viên những kỹ năng cần thiết để nghiên cứu và áp dụng các công nghệ xử lý ngôn ngữ tự nhiên vào thực tiễn. Việc hiểu và áp dụng những kiến thức từ môn học này không chỉ giúp sinh viên nắm bắt được xu hướng công nghệ hiện đại mà còn đem lại nền tảng vững chắc cho sự nghiệp nghiên cứu và phát triển trong tương lai.

2. **Giới thiệu về bài toán nhận diện cảm xúc qua giọng nói**

   Bài toán nhận diện cảm xúc qua giọng nói là một trong những thách thức quan trọng trong lĩnh vực xử lý thông tin giọng nói. Đây là một lĩnh vực nghiên cứu tập trung vào việc phân tích và xác định các cảm xúc (như vui vẻ, buồn bã, bực bội, sợ hãi,…) từ các tín hiệu giọng nói của con người.

   Việc nhận diện cảm xúc qua giọng nói mang lại nhiều ứng dụng trong thực tế, bao gồm:

   - Công nghệ trợ lý ảo: Giúp trợ lý ảo nhận biết tâm trạng của người dùng để tương tác hiệu quả hơn.
   - Giám sát sức khỏe tâm lý: Dùng để đánh giá tâm trạng của cá nhân trong các bệnh lý như trầm cảm, lo âu.
   - Giao tiếp giữa người và máy móc: Cải thiện khả năng tương tác giữa con người và các hệ thống tự động.

   Bài toán này đòi hỏi sự kết hợp của nhiều kỹ thuật xử lý tín hiệu âm thanh và học máy. Các đặc điểm của giọng nói như tốc độ, âm sắc, cường độ, và các đặc trưng ngữ điệu phải được trích xuất và sử dụng để huấn luyện các mô hình máy học nhằm phân loại và nhận diện cảm xúc.

   Đây là một bài toán có tính đa dạng và đòi hỏi sự nghiên cứu sâu sắc để giải quyết các thách thức như biến đổi của giọng nói trong các điều kiện khác nhau, sự khác biệt về văn hóa và ngôn ngữ, cũng như tính độc đáo của từng cá nhân. Việc thành công trong việc nhận diện cảm xúc qua giọng nói có thể đem lại những ứng dụng có giá trị cao trong thực tế và tạo ra những tiến bộ đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo.

Chương 2: NHẬN DIỆN TÌNH HUỐN ĐỐI THOẠI TRONG CUỘC TRÒ CHUYỆN HẰNG NGÀY BẰNG THÔNG TIN ĐA PHƯƠNG TIỆN

1. Giới thiệu chung về bài báo

Bài báo "Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information" của Yuya Chiba và Ryuichiro Higashinaka là một nghiên cứu quan trọng trong lĩnh vực nhận diện tình huống hội thoại bằng cách sử dụng thông tin đa phương thức. Tình huống hội thoại trong bài nghiên cứu này bao gồm nhiều yếu tố như phong cách hội thoại, địa điểm, hoạt động và mối quan hệ giữa các người tham gia. Đây là một nghiên cứu tiên phong trong việc sử dụng các phương pháp học máy hiện đại để nhận diện và phân tích các tình huống hội thoại trong đời sống hàng ngày, với mục tiêu cuối cùng là phát triển các hệ thống hội thoại thông minh có khả năng thích ứng với nhiều tình huống khác nhau.

2. Nội dung nghiên cứu

Bài báo bắt đầu bằng việc giới thiệu tổng quan về nhận diện tình huống hội thoại và sự cần thiết của việc sử dụng thông tin đa phương thức để cải thiện độ chính xác của các mô hình nhận diện. Thông tin đa phương thức ở đây bao gồm âm thanh, hình ảnh và ngôn ngữ, tất cả đều được sử dụng để tạo ra một bức tranh toàn diện về tình huống hội thoại. Mục tiêu của nghiên cứu là xây dựng một mô hình có khả năng sử dụng đồng thời cả ba loại thông tin này để nhận diện các tình huống hội thoại một cách chính xác.

3. Phương pháp

Để đạt được mục tiêu này, nghiên cứu đã sử dụng bộ dữ liệu CEJC (Corpus of Everyday Japanese Conversation), một tập dữ liệu phong phú ghi lại các cuộc hội thoại tự nhiên trong cuộc sống hàng ngày. Bộ dữ liệu này bao gồm các đoạn video, âm thanh và văn bản của các cuộc hội thoại giữa 40 người tham gia. Các cuộc hội thoại được ghi lại bằng nhiều thiết bị ghi âm và camera khác nhau, tạo ra một nguồn dữ liệu đa phương thức phong phú và đa dạng.

Một phần quan trọng của nghiên cứu là việc xây dựng mô hình nhận diện tình huống hội thoại. Mô hình này bao gồm các bước chính sau:

1.1. Chuyển đổi đầu vào

Các thông tin đa phương thức từ dữ liệu đầu vào được chuyển đổi thành các vector biểu diễn. Đối với hình ảnh, mô hình sử dụng ResNet50 để trích xuất đặc trưng. Đối với âm thanh, VGGish được sử dụng để trích xuất đặc trưng âm thanh. Đối với ngôn ngữ, BERT được sử dụng để trích xuất đặc trưng ngôn ngữ. Các vector biểu diễn này sau đó được đưa vào các mạng Uni-GRUs để tạo ra các biểu diễn cuối cùng của từng loại thông tin.

1.2. Kết hợp và dự đoán

Các vector biểu diễn cuối cùng từ các Uni-GRUs sau đó được kết hợp lại và đưa vào các lớp đầu ra để dự đoán kết quả nhận diện tình huống hội thoại. Mô

hình sử dụng phương pháp học đa nhiệm để huấn luyện, nhằm tối ưu hóa việc nhận diện các tình huống hội thoại khác nhau cùng một lúc.

4. Kết quả

Kết quả của nghiên cứu cho thấy mô hình đề xuất đã đạt được hiệu suất cao trong việc nhận diện các tình huống hội thoại. Mô hình được đánh giá trên các tập dữ liệu kiểm tra với các chỉ số như độ chính xác, độ nhạy và F1-score. Kết quả cho thấy mô hình có thể nhận diện các tình huống hội thoại với độ chính xác cao hơn so với mức ngẫu nhiên, cho thấy tính khả thi của việc sử dụng thông tin đa phương thức để cải thiện độ chính xác của các mô hình nhận diện tình huống hội thoại.

Một phần quan trọng khác của nghiên cứu là so sánh kết quả nhận diện của mô hình với khả năng nhận diện của con người. Các nhà nghiên cứu đã thực hiện một thí nghiệm để đánh giá khả năng nhận diện tình huống hội thoại bằng con người và sử dụng kết quả này làm tham chiếu để so sánh với mô hình. Kết quả cho thấy mặc dù mô hình vẫn còn khoảng cách so với khả năng nhận diện của con người, nhưng đã đạt được kết quả khá tốt, cho thấy tiềm năng của các phương pháp học máy trong việc nhận diện tình huống hội thoại.

5. Phân tích

Nghiên cứu đã chứng minh được tính hiệu quả của việc sử dụng thông tin đa phương thức trong nhận diện tình huống hội thoại. Việc sử dụng các đặc trưng hình ảnh, âm thanh và ngôn ngữ đã giúp cải thiện độ chính xác của mô hình, so với việc chỉ sử dụng một loại thông tin duy nhất. Điều này cho thấy rằng các tình huống hội thoại trong đời sống hàng ngày thường phức tạp và đa chiều, đòi hỏi các mô hình nhận diện phải có khả năng xử lý và kết hợp nhiều loại thông tin khác nhau.

Một điểm mạnh của nghiên cứu là việc sử dụng phương pháp học đa nhiệm để huấn luyện mô hình. Phương pháp này không chỉ giúp tối ưu hóa việc nhận diện các tình huống hội thoại khác nhau cùng một lúc, mà còn giúp cải thiện hiệu suất tổng thể của mô hình. Phương pháp học đa nhiệm cho phép mô hình học được các mối quan hệ và sự phụ thuộc giữa các tình huống hội thoại khác nhau, từ đó cải thiện độ chính xác và độ nhạy của các dự đoán.

Tuy nhiên, nghiên cứu cũng gặp phải một số thách thức và hạn chế. Một trong những thách thức lớn nhất là sự phức tạp và đa dạng của các tình huống hội thoại trong đời sống hàng ngày. Mặc dù bộ dữ liệu CEJC đã cung cấp một nguồn dữ liệu phong phú và đa dạng, nhưng vẫn có những tình huống hội thoại đặc thù mà mô hình không thể nhận diện chính xác. Điều này cho thấy cần có thêm các nghiên cứu và cải tiến để mô hình có thể xử lý được tất cả các tình huống hội thoại khác nhau.

Ngoài ra, việc so sánh kết quả nhận diện của mô hình với khả năng nhận diện của con người cũng cho thấy một khoảng cách đáng kể. Mặc dù mô hình đã đạt được kết quả khá tốt, nhưng vẫn chưa thể vượt qua được khả năng nhận diện

của con người. Điều này cho thấy cần có thêm các nghiên cứu để cải thiện mô hình, nhằm giảm khoảng cách này và đạt được kết quả tốt hơn.

6. Kết luận

Nghiên cứu "Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information" đã đưa ra một phương pháp hiệu quả để nhận diện tình huống hội thoại bằng cách sử dụng thông tin đa phương thức. Kết quả của nghiên cứu cho thấy mô hình đề xuất có thể nhận diện các tình huống hội thoại với độ chính xác cao hơn so với mức ngẫu nhiên và tiềm năng của việc sử dụng các phương pháp học máy hiện đại trong việc phát triển các hệ thống hội thoại thông minh.

Tuy nhiên, nghiên cứu cũng chỉ ra rằng vẫn còn nhiều thách thức và hạn chế cần được giải quyết. Cần có thêm các nghiên cứu và cải tiến để mô hình có thể xử lý được tất cả các tình huống hội thoại khác nhau trong đời sống hàng ngày và giảm khoảng cách với khả năng nhận diện của con người. Điều này sẽ mở ra nhiều hướng nghiên cứu mới và đóng góp quan trọng vào sự phát triển của các hệ thống hội thoại thông minh trong tương lai.

Chương 3: NHẬN BIẾT CẢM XÚC CỦA GIỌNG NÓI

1. Tóm tắt

Nhận diện cảm xúc giọng nói là một lĩnh vực nghiên cứu thu hút sự quan tâm ngày càng tăng trong trí tuệ nhân tạo. Nó tập trung vào việc phát triển các hệ thống có thể tự động nhận diện và phân loại cảm xúc được thể hiện qua giọng nói của con người. Khả năng này có tiềm năng ứng dụng rộng rãi trong nhiều lĩnh vực như dịch vụ khách hàng, giáo dục, chăm sóc sức khỏe, và giải trí.

Bài báo này trình bày nghiên cứu về việc nhận diện cảm xúc giọng nói sử dụng hai mô hình học máy phổ biến: Random Forest và K-Nearest Neighbors (KNN). Thực nghiệm được thực hiện trên hai bộ dữ liệu lớn: RAVDESS và TESS, bao gồm bản ghi âm giọng nói của nhiều người thể hiện đa dạng cảm xúc.

Điểm mạnh của hướng đi này so với các phương pháp cũ đó là không bị giới hạn bởi bộ dữ liệu, hay nói cách khác, mặc dù nghiên cứu này chỉ sử dụng 2 bộ dữ liệu tiếng Anh nhưng có thể hoạt động tốt với các âm thanh của ngôn ngữ khác. Lý giải cho việc có sự vượt trội này là vì nghiên cứu của chúng tôi sử dụng đặc trưng MFCCs thay cho phương án trích xuất văn bản từ âm thanh, nhờ đó có thể đưa ra dự đoán cảm xúc dựa trên âm điệu của giọng nói (tương tự như cách tai người phát hiện cảm xúc trong giao tiếp)

Kết quả cho thấy mô hình KNN đạt hiệu quả cao hơn Random Forest trong việc phân loại cảm xúc giọng nói, với độ chính xác đạt 95,2% so với 85,5%. Khả năng nhận diện cảm xúc giọng nói chính xác có tiềm năng ứng dụng rộng rãi trong nhiều lĩnh vực như dịch vụ khách hàng, giáo dục, chăm sóc sức khỏe, và giải trí.

2. Giới thiệu

a. Vai trò của cảm xúc giọng nói trong giao tiếp

Cảm xúc là một phần thiết yếu của con người, ảnh hưởng đến mọi khía

cạnh trong cuộc sống. Trong giao tiếp, cảm xúc được thể hiện qua nhiều kênh khác nhau, bao gồm ngôn ngữ, biểu cảm khuôn mặt và giọng nói. Nhận biết được cảm xúc của người đối diện giúp chúng ta:

- Hiểu rõ hơn về ý định và thông điệp của họ: Cảm xúc có thể bổ sung hoặc mâu thuẫn với nội dung lời nói, giúp chúng ta hiểu rõ hơn về ý định thực sự của người nói.
- Đánh giá mức độ tin cậy: Giọng nói có thể tiết lộ những cảm xúc mà người nói cố gắng che giấu, giúp chúng ta đánh giá mức độ tin cậy của thông tin được truyền tải.
- Tăng cường sự đồng cảm và kết nối: Khi chúng ta hiểu được cảm xúc của người đối diện, chúng ta có thể dễ dàng đồng cảm và kết nối với họ hơn, từ đó xây dựng mối quan hệ tốt đẹp.

b. Nhận diện cảm xúc giọng nói

Nhận diện cảm xúc giọng nói (Speech Emotion Recognition - SER) là một lĩnh vực nghiên cứu thuộc trí tuệ nhân tạo, tập trung vào việc phát triển các hệ thống có thể tự động nhận diện và phân loại cảm xúc được thể hiện qua giọng nói của con người. SER có nhiều ứng dụng tiềm năng trong nhiều lĩnh vực như:

- **Dịch vụ khách hàng:** Giúp các hệ thống tự động đánh giá mức độ hài lòng của khách hàng và cung cấp dịch vụ phù hợp hơn.
- **Giáo dục:** Giúp giáo viên nhận biết những học sinh đang gặp khó khăn hoặc cần được hỗ trợ.
- **Giải trí:** Giúp các hệ thống giải trí tương tác với người dùng một cách tự nhiên và hấp dẫn hơn.

3. Phương pháp tiếp cận

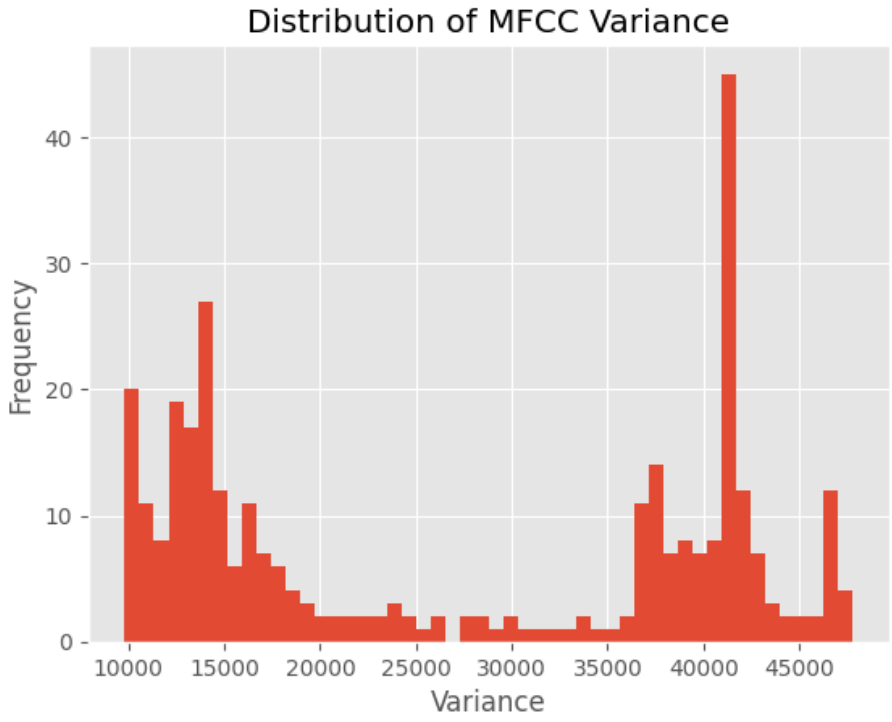
a. Chia khung

Đặc điểm của dữ liệu: Qua thực nghiệm cho thấy bộ dữ liệu có tần số âm thanh dao động trong khoảng 16000 đến 24000 Hz. Dải tần số này bao gồm hầu hết các thành phần âm thanh có thể nghe được bởi con người và đóng vai trò quan trọng trong việc truyền tải cảm xúc qua giọng nói. Tín hiệu âm thanh được chia thành các khung ngắn (frame), mỗi khung có độ dài khoảng 500ms. Việc chia frame giúp phân tích từng phần nhỏ của tín hiệu âm thanh và trích xuất các đặc trưng quan trọng liên quan đến cảm xúc. Bên cạnh đó là để có thể tránh được các khoảng dữ liệu nhiễu hoặc không có nhiều ý nghĩa trong phân tích.

b. Trích xuất đặc trưng

Sau khi chia tệp âm thanh thành các frame, các đặc trưng của âm thanh đươc trích xuất từ mỗi khung dữ liệu. Ở bài báo này, nhóm chúng tôi đã sử dụng Mel-frequency cepstral coefficients (MFCC) là một tập hợp các đặc trưng được tính toán từ phổ âm thanh. MFCC mô tả sự phân bô năng lượng âm thanh trên các dải tần số mel, mô phỏng các thức mà tai người cảm nhận âm thanh. Để thực hiện loại bỏ các khung có ít ý nghĩa hơn, nhóm chúng tôi đã thực hiện tính toán phương sai MFCC của từng khung,

sau đó loại bỏ các khung có kết quả thuộc lớp có tầng suất xuất hiện thấp hơn.



1Phân phối tần suất giá trị phương sai của MFCC (trong 1 khung)

c. Dán nhãn dữ liệu

Sau khi đã thực hiện việc sàn lọc các MFCC, chúng tôi đi đến phần dán nhãn dữ liệu. Trong thực nghiệm này, chúng tôi đưa ra 7 nhãn dữ liệu khác nhau tương ứng với 7 loại cảm xúc cơ bản. Nhãn cảm xúc ban đầu của các đoạn MFCC là chuỗi ký tự (“happy”, “sad”,…), do đó, để mô hình học máy có thể xử lý ở các bước tiếp theo, chúng tôi chuyển đổi các nhãn này thành các giá trị số thông qua mã hóa nhãn (label encoding). Sau đó, nhóm chúng tôi đã tiếp tục chuyển đổi các giá trị số này thành các cột biến giả (one-hot encoding) để phù hợp cho các mô hình học máy mà chúng tôi sẽ tiếp cận ở phần sau. One-hot encoding tạo ra một vector nhị phân với độ dài bằng số lượng nhãn cảm xúc, trong đó chỉ có một phần tử có giá trị 1 (đại diện cho nhãn cảm xúc tương ứng) và các phần tử còn lại có giá trị 0.

4. Bộ dữ liệu

Trong nghiên cứu này, chúng tôi đã sử dụng 2 bộ dữ liệu âm thanh là RAVDESS và TESS.

RAVDESS (Real-time Audio-Visual Database of Emotion Expressions), là một bộ dữ liệu đa phương tiện bao gồm nhiều bản ghi âm giọng nói và video của 8 người nói thể hiện 8 cảm xúc cơ bản là vui, buồn, tức giận, sợ hãi, ngạc nhiên, ghê tởm, trung lập và bình tĩnh. Ở trong bài nghiên cứu này, do giới hạn về tài nguyên, chúng tôi đã xếp cảm xúc ‘trung lập’ và ‘bình tĩnh’ vào cùng một lớp là ‘trung lập), bên cạnh đó, chúng tôi cũng chỉ sử dụng phần dữ liệu về âm thanh của bộ dữ liệu (không bao gồm video).

TESS (Toronto Emotion Speech Synthesis) là một bộ dữ liệu âm thanh bao gồm bản ghi âm giọng nói của 13 người nói thể hiện 7 cảm xúc, tương tự như bộ dữ liệu RAVDESS (trừ cảm xúc ‘bình tĩnh’).

Hai bộ dữ liệu này có kích thước tương đối lớn, với 2800 file audio và 1440 file audio ở bộ TESS và RAVDESS. Mỗi tệp âm thanh có độ dài từ 3 đến 5 giây, sau các bước tiền xử lý vừa được đề cập ở trên sẽ cho ra một bộ dữ liệu các giá trị MFCC lớn (khoảng hơn 220000 dòng và 20 cột, chưa bao gồm cột nhãn cảm xúc)

5. **Mô hình huấn luyện**

Sau khi dữ liệu đã được tiền xử lý và sẵn sàng, bước tiếp theo là huấn luyện mô hình học máy để nhận diện cảm xúc giọng nói. Trong nghiên cứu này, chúng tôi đã sử dụng bốn mô hình học máy là Random Forest, K-Nearest Neighbors (KNN), LSTM, và SVM để tiếp cận bài toán.

Tại sao không sử dụng BERT?

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình học sâu mạnh mẽ chủ yếu được thiết kế cho các tác vụ xử lý ngôn ngữ tự nhiên (NLP) như phân loại văn bản, trả lời câu hỏi và các nhiệm vụ khác liên quan đến văn bản. Tuy nhiên, bài toán nhận diện cảm xúc giọng nói của chúng tôi dựa trên việc xử lý đặc trưng âm thanh (MFCCs - Mel-frequency cepstral coefficients) chứ không phải chuyển tín hiệu âm thanh sang văn bản rồi mới dự đoán cảm xúc. MFCCs là các đặc trưng phổ biến được sử dụng để mô tả các thuộc tính tần số của tín hiệu âm thanh, giúp phân biệt các đặc điểm khác nhau của giọng nói.

Do BERT không được tối ưu hóa cho việc xử lý trực tiếp các tín hiệu âm thanh hay các đặc trưng như MFCCs, nên việc sử dụng các mô hình như Random Forest, KNN, LSTM, và SVM là phù hợp hơn. Những mô hình này đã được chứng minh hiệu quả trong việc xử lý và phân loại các đặc trưng âm thanh, giúp chúng tôi đạt được kết quả tốt trong bài toán nhận diện cảm xúc giọng nói.

a. **Mô hình Random Forest**

Random Forest là một mô hình học máy mạnh mẽ dựa trên phương pháp cây quyết định. Nó hoạt động bằng cách tạo ra nhiều cây quyết định và tổng hợp kết quả của chúng để đưa ra dự đoán cuối cùng. Phương pháp này giúp giảm thiểu vấn đề overfitting và cải thiện độ chính xác của mô hình.

Để thiết lập mô hình Random Forest, chúng tôi sử dụng thư viện scikit-learn, một thư viện phổ biến cho học máy trong Python. Chúng tôi điều chỉnh các tham số quan trọng như số lượng cây (n_estimators), độ sâu tối đa của cây (max_depth) và các tham số khác để tối ưu hóa hiệu quả của mô hình. Kết quả thực nghiệm cho thấy số lượng cây tối ưu cho bài toán này là 200, và độ sâu tối đa tốt nhất là 16.

Sử dụng kỹ thuật Randomized Search Cross-Validation, chúng tôi tìm kiếm các tham số tối ưu cho mô hình. Randomized Search giúp thử

nghiệm nhiều tổ hợp tham số một cách hiệu quả mà không cần phải kiểm tra tất cả các tổ hợp có thể, tiết kiệm thời gian và tài nguyên tính toán.
Sau khi huấn luyện, mô hình Random Forest đạt được độ chính xác 85.5%, cho thấy khả năng phân loại cảm xúc giọng nói khá tốt trên bộ dữ liệu này.

b. Mô hình K-Nearest Neighbors (KNN)

KNN là một mô hình học máy đơn giản nhưng hiệu quả, dựa trên nguyên lý "gần kề nhất". Nó phân loại một mẫu dữ liệu mới dựa trên nhãn của các mẫu dữ liệu gần nhất trong không gian đặc trưng. Mô hình này không yêu cầu quá trình huấn luyện phức tạp và có thể dễ dàng triển khai.
Tương tự như Random Forest, chúng tôi sử dụng scikit-learn để thiết lập và huấn luyện mô hình KNN. Các tham số quan trọng như số lượng láng giềng (n_neighbors), trọng số (weights) và thuật toán tìm kiếm (algorithm) được điều chỉnh để tối ưu hóa hiệu quả của mô hình. Qua thực nghiệm cho thấy số lượng "láng giềng" tối ưu nhất cho trường hợp bộ dữ liệu này là 3.
Để tìm các tham số tối ưu cho KNN, chúng tôi sử dụng Grid Search Cross-Validation, một phương pháp tìm kiếm toàn diện bằng cách kiểm tra tất cả các tổ hợp tham số có thể. Mặc dù tốn nhiều thời gian hơn Randomized Search, Grid Search đảm bảo rằng chúng tôi tìm được tổ hợp tham số tốt nhất cho mô hình.
Kết quả huấn luyện cho thấy mô hình KNN đạt độ chính xác 95.2%, cao hơn so với Random Forest. Điều này cho thấy KNN có khả năng phân loại cảm xúc giọng nói tốt hơn trên bộ dữ liệu này.

c. Mô hình LSTM (Long Short-Term Memory)

Mô hình LSTM (Long Short-Term Memory) được áp dụng trong nghiên cứu là một kiến trúc mạng nơ-ron hồi quy sâu, được thiết kế đặc biệt để xử lý và dự đoán các chuỗi thời gian phức tạp, như các đặc trưng âm thanh như MFCCs (Mel-frequency cepstral coefficients). Mô hình của chúng tôi bao gồm ba lớp LSTM với số đơn vị nơ-ron là 500, được sắp xếp theo thứ tự từ đầu vào đến đầu ra của chuỗi dữ liệu. Các lớp LSTM được kết nối với các lớp Dropout với tỷ lệ 0.1 để hạn chế hiện tượng overfitting trong quá trình huấn luyện.
Sau khi thiết lập mô hình, chúng tôi sử dụng thư viện Keras trong Python để huấn luyện mô hình LSTM trên tập dữ liệu. Qua quá trình huấn luyện, mô hình đạt được độ chính xác trên tập kiểm tra là 81%.

d. Mô hình SVM (Support Vector Machine)

Mô hình SVM (Support Vector Machine) được áp dụng trong nghiên cứu là một trong những phương pháp học máy phổ biến và mạnh mẽ, được sử dụng để phân loại và dự đoán dữ liệu. Chúng tôi sử dụng kernel tuyến tính (linear kernel) cho mô hình SVM này, cho phép mô hình xây dựng ranh giới phân chia tuyến tính giữa các lớp dữ liệu cảm xúc từ giọng nói.
Để thiết lập mô hình, chúng tôi sử dụng thư viện scikit-learn trong Python và cấu hình mô hình với khả năng tính toán xác suất (probability=True) để thu thập thông tin xác suất dự đoán. Mô hình được huấn luyện trên tập dữ liệu đã được tiền xử lý với các đặc trưng MFCCs (Mel-frequency cepstral coefficients). Quá trình huấn luyện mô hình SVM với kernel tuyến tính giúp tối ưu hóa khả năng phân loại và tránh hiện tượng overfitting. Kết quả thực nghiệm cho thấy mô hình SVM đạt được độ chính xác trên tập kiểm tra là 77%.

e. Một số điểm nổi bật trong quá trình thiết lập mô hình

- Chọn lựa tham số kỹ càng: Cả hai mô hình đều được tối ưu hóa bằng cách thử nghiệm nhiều tổ hợp tham số khác nhau thông qua các kỹ thuật tìm kiếm như Grid Search và Randomized Search. Việc chọn lựa tham số kỹ càng giúp cải thiện đáng kể hiệu quả của mô hình.
- Xử lý và tiền xử lý dữ liệu: Việc tiền xử lý dữ liệu cẩn thận, bao gồm trích xuất đặc trưng MFCC và chuyển đổi nhãn, đóng vai trò quan trọng trong việc nâng cao độ chính xác của mô hình. Đặc biệt, việc sử dụng các đặc trưng âm thanh chất lượng cao giúp mô hình học được các đặc tính quan trọng từ dữ liệu.
- Chia tách dữ liệu: Việc chia tách dữ liệu thành tập huấn luyện và tập kiểm tra đảm bảo rằng mô hình được đánh giá chính xác về khả năng tổng quát hóa và không bị overfitting. Chúng tôi sử dụng tỷ lệ chia tách 70/30 để đảm bảo có đủ dữ liệu cho cả hai quá trình.
- Sử dụng Cross-Validation: Sử dụng kỹ thuật Cross-Validation giúp đánh giá độ chính xác của mô hình trên nhiều tập dữ liệu khác nhau, giảm thiểu nguy cơ overfitting và đảm bảo rằng mô hình có hiệu quả trên dữ liệu chưa thấy trước.

6. Kết quả

Mô hình	Độ chính xác (Accuracy)
Random Forest	85.5%
KNN	95.2%
LSTM	81%
SVM	77%

Dựa trên kết quả độ chính xác của các mô hình học máy, chúng ta nhận thấy KNN đạt độ chính xác cao nhất với 95.2%, tiếp đó là Random Forest với 85.5%. LSTM và SVM có kết quả lần lượt là 81% và 77%. KNN và Random Forest thể hiện khả năng phân loại cảm xúc từ giọng nói tốt hơn so với LSTM và SVM trong bài toán này.

Chương 4: KẾT LUẬN

Nhận diện cảm xúc giọng nói (SER) là một lĩnh vực nghiên cứu tiềm năng với nhiều ứng dụng trong thực tiễn. Trong dịch vụ khách hàng, hệ thống SER có thể đánh giá tâm trạng khách hàng theo thời gian thực, giúp nhân viên điều chỉnh giao tiếp để tăng sự hài lòng. Các trợ lý ảo như Siri, Google Assistant có thể trở nên thông minh hơn khi hiểu được cảm xúc của người dùng, giúp hỗ trợ hiệu quả hơn.

Trong giáo dục, hệ thống SER có thể giúp giáo viên nhận diện cảm xúc của học sinh trong học trực tuyến, phát hiện sớm các vấn đề học tập và tâm lý. Trong chăm sóc sức khỏe, SER có thể theo dõi tâm trạng và cảm xúc của bệnh nhân, cung cấp dữ liệu hữu ích cho chuyên gia y tế. Trong an ninh, SER có thể phát hiện cảm xúc như sợ hãi hoặc căng thẳng từ giọng nói, hỗ trợ điều tra và ngăn chặn hành vi nguy hiểm.

Hướng phát triển của SER bao gồm cải thiện độ chính xác và độ tin cậy bằng cách sử dụng các mô hình học sâu tiên tiến như CNN, RNN, và Transformer. Nghiên cứu cũng tập trung vào phát triển hệ thống đa ngữ, tích hợp với các cảm biến khác như hình ảnh từ camera, và dữ liệu từ thiết bị đeo tay. Việc đưa các hệ thống SER vào ứng dụng thực tế giúp kiểm tra khả năng hoạt động trong các điều kiện thực, đồng thời phát triển giao diện người dùng thân thiện để dễ dàng tương tác và nhận lợi ích tối đa từ công nghệ này. Nhận diện cảm xúc giọng nói không chỉ hứa hẹn cải thiện cuộc sống hàng ngày mà còn nâng cao hiệu quả công việc và chất lượng cuộc sống.

TÀI LIỆU THAM KHẢO

Chiba, Yuya, and Ryuichiro Higashinaka. "**Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information.**" *Interspeech*. 2021.

Ingale, A. B., & Chaudhari, D. S. (2012). **Speech emotion recognition**. *International Journal of Soft Computing and Engineering (IJSCE), 2*(1), 235-238.

“Speech Emotion Regconition” Repository

RAVDESS Emotional speech audio (kaggle.com)

Toronto emotional speech set (TESS) | TSpace Repository (utoronto.ca)

PHỤ LỤC