

BÁO CÁO CUỐI KỲ

Môn học

CS2205.CH1501 -

PHƯƠNG PHÁP LUẬN NCKH

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

03/2021 - 06/2021

----- *Trang này cố tình để trống* -----

HƯỚNG DẪN

Yêu cầu:


- *Bước 1: Chọn File/Make a copy để tạo ra một file theo template mẫu https://docs.google.com/document/d/1pu86lH6STGaVk2JH70n3jWx8qt9Eue_imVTQhg3s/. Đặt tên tập tin này là: CS2205.CH1501.RM.FinalReport.MSHV*
- *Bước 2: Điền các thông tin về đề cương đề tài vào file GDocs trên. Tối đa 6 trang.*
- *Bước 3: Copy toàn bộ nội dung đề cương đề tài và Paste vào cuối tập tin này (tránh ghi đè lên nội dung của HV khác).*
- *Bước 4: Nộp bài (Turn in) theo yêu cầu trên Classroom. Chọn Add or Create và chọn Link đến file Google ở trên. Lưu ý đặt quyền Anyone with the link - Viewer. Trong phần Private Comment, cung cấp thông tin của github repos, thông tin các thành viên của nhóm và các ghi chú khác nếu có. Lưu một phiên bản pdf của đề cương trên github repos*

Lưu ý:

- *Việc tuân thủ các hướng dẫn, các yêu cầu theo mẫu là bắt buộc và được đánh giá trong điểm tổng kết của đồ án môn học.*
- ***Deadline: 25/07/2021***

----- *Trang này cố tình để trống* -----

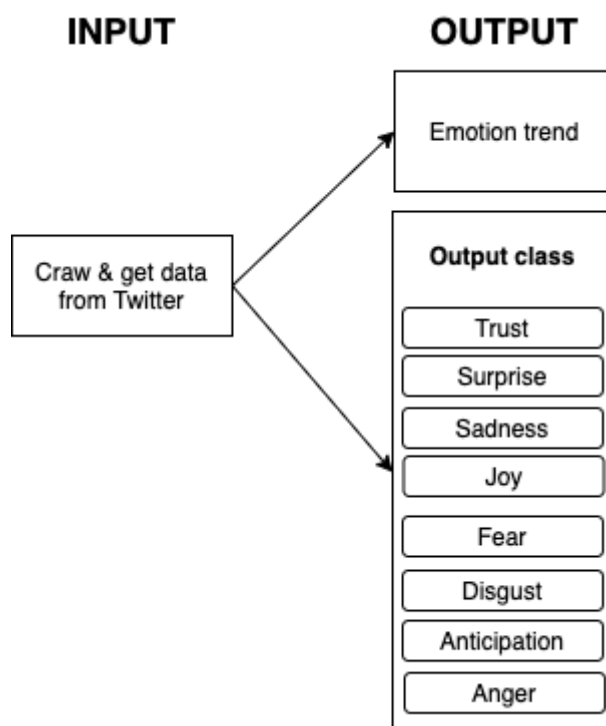
ĐỀ CƯƠNG LUẬN VĂN THẠC SĨ

Họ và tên (IN HOA)	NGUYỄN VĂN VIẾT
Ảnh	
Số buổi vắng	1
Bonus	31 lần comment trên Google Classroom
Tên đề tài (VN)	PHÁT HIỆN VÀ PHÂN TÍCH XU HƯỚNG CẢM XÚC CỦA NGƯỜI TRÊN MẠNG XÃ HỘI TRONG GIAI ĐOẠN GIÃN CÁCH XÃ HỘI
Tên đề tài (EN)	
Giới thiệu	<ul style="list-style-type: none"> ● <i>Bài toán/vấn đề mà đề tài muốn giải quyết</i> <ul style="list-style-type: none"> - Phát hiện và phân tích xu hướng cảm xúc của người trên mạng xã hội trong giai đoạn giãn cách xã hội ● <i>Lý do chọn đề tài, khả năng ứng dụng thực tế, tính thời sự</i> <ul style="list-style-type: none"> - Trước tình hình diễn biến phức tạp của dịch COVID, các quốc gia đều đưa ra các quy định chỉ thị về giãn cách xã hội với nhiều quy mô khác

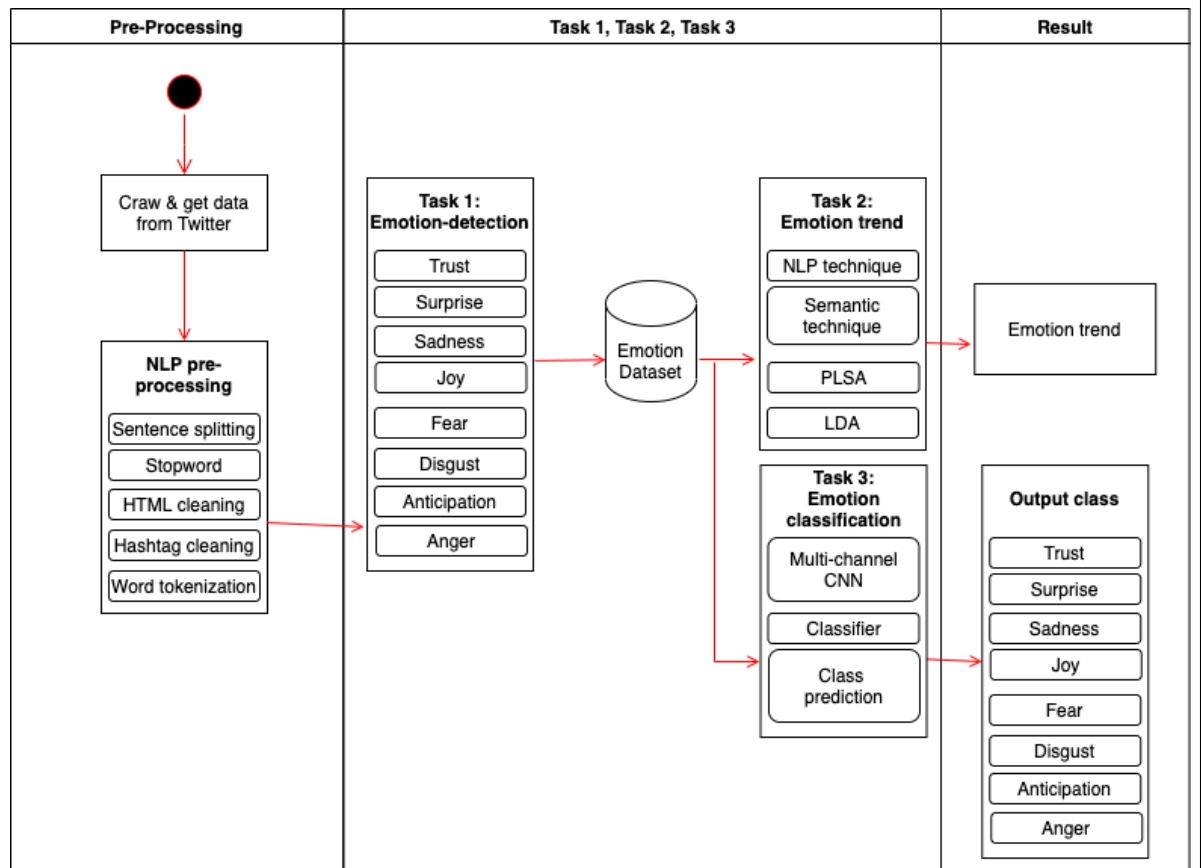
nhau nhằm hạn chế sự lây lan chéo. Điều này tạo ra nhiều xu hướng cảm xúc tới xã hội, bằng cách sử dụng mạng xã hội, nhiều người dùng đã bày tỏ cảm xúc của bản thân trong giai đoạn giãn cách này. Để đánh giá phản ứng của xã hội đối với giai đoạn giãn cách các nhà phân tích cần một công cụ nhằm phát hiện và đánh giá cảm xúc của người trong đó có việc đánh giá thông qua các bài đăng trên mạng xã hội Twitter. Với công cụ này, các nhà đánh giá sẽ có cái nhìn tức thời và mang tính thời sự về cảm nhận của người dân nhằm đưa ra các chính sách, chỉ thị hợp lý hơn trong thời gian tới.

- *Mô tả input và output*

- Input là bộ dataset gồm các bài đăng (post) bài đăng có hashtag (#stayathome, #coronavirus, #covid-19, #quarantine, #lockdown, #social-distancing)
- Output là dựa vào các mô hình máy học để phát triển 1 bộ phát hiện cảm xúc và xu hướng cảm xúc.



Mục tiêu	<ul style="list-style-type: none"> • Mục tiêu của luận văn này là đóng góp một phương pháp luận dựa trên AI để có thể phân tích các xu hướng cảm xúc để hiểu rõ hơn về tác động của các quy định hay chỉ thị về kiểm dịch và giãn cách xã hội. Hai mục tiêu chính là: 1) xây dựng và phát triển mô hình học máy để phát hiện cảm xúc, 2) thí điểm mô hình này trên các tweet không có cấu trúc của mạng xã hội Twitter trong thời gian giãn cách xã hội. Các mục tiêu trên được đánh giá thông qua các đóng góp sau: Xây dựng 1 bộ dataset bao gồm các tweet về COVID-19 đã được emotion-annotated, đây là cơ sở để xây dựng và triển khai các hệ thống phát hiện cảm xúc dựa trên mô hình máy học trong tương lai; Thiết kế multi-task framework để phân tích cảm xúc trên 8 vị trí tiêu chuẩn qua mô hình Plutchik (Emotion classification); Khám phá các xu hướng ngữ nghĩa từ thông qua các mô hình khác nhau như LDA (latent Dirichlet allocation) hay PLSA (probabilistic latent semantic analysis) dựa trên các tweets; và sử dụng mô hình convolutional neural network (CNN) cho phát hiện cảm xúc từ COVID-19 tweets.
Nội dung và phương pháp thực hiện	<ul style="list-style-type: none"> - Nature language processing (NLP) và machine learning (ML) được sử dụng trong bối cảnh xác định loại cảm xúc trong tweet text. Luận văn tập trung vào phát hiện cảm xúc trong sức khỏe cộng đồng trực tuyến, mô hình Lexical dựa trên cảm xúc, và hướng đưa ra quyết định các chính sách chỉ thị bảo vệ sức khỏe cộng đồng sử dụng phân tích tweets liên quan đến COVID-19. - Các bước thực hiện bao gồm: xử lý dữ liệu văn bản đầu vào, phát hiện cảm xúc và đánh giá điểm cường độ, tính toán xu hướng cảm xúc và cuối cùng là đánh giá thuật toán học sâu sử dụng training và testing data.



- Dựa trên mô hình Plutchik emotion và kỹ thuật học sâu để xây dựng khung làm việc đa tác (multitask framework) từ đó giải quyết các mục tiêu nghiên cứu của luận văn. Các tiếp cận bao gồm 3 công việc (task) chính.
 - + Pre-processing: Từ dataset gốc là những tweet có hashtag (#stayathome, #coronavirus, #covid-19, #quarantine, #lockdown, #social-distancing), tiến hành tiền xử lý bằng các kỹ thuật tách câu (sentence splitting), tách từ (word tokenization), loại bỏ stop-words và hashtags, và làm sạch HTML (nếu có) để làm sạch, loại bỏ nhiễu và đánh giá mức độ liên quan tới chủ đề.
 - + Task 1 Emotion-detection: Xây dựng emotional vector cho tweet bằng Plutchik emotion Lexicon, sau đó đánh giá và ghi nhận emotion có điểm cao nhất trong vector trên là cảm xúc của tweet đó.

	<ul style="list-style-type: none"> + Task 2 xu hướng cảm xúc (emotion) và ngữ nghĩa (semantic): Việc hiểu biết các thay đổi trạng thái cảm xúc theo thời gian được xem là 1 yếu tố quan trọng ảnh hưởng đến quyết định thay đổi chính sách về giãn cách xã hội và sức khỏe cộng đồng. Để đạt được ý nghĩa này, mô hình PLSA và LDA được ứng dụng. Mô hình PLSA được sử dụng như một kỹ thuật NLP để khai phá sự tương đồng về chủ đề giữa các từ. Trong khi đó mô hình LDA rất hữu dụng để chiết xuất ngữ nghĩa và tạo xu hướng dựa trên thời gian về mặt ngữ nghĩa. + Task 3 tạo lập mô hình câu và phát hiện COVID-19 emotion: sử dụng mô hình CNN to triển khai một hệ thống emotion detection dựa trên emotion vectors. Các layer của mô hình CNN này bao gồm: embedding layers, convolution layers, drop out layers, Max-pooling layer.
Kết quả dự kiến	<ul style="list-style-type: none"> ● <i>Phần mềm ứng dụng</i> - Ứng dụng trong công cụ phân tích cảm xúc cộng đồng trước và sau các quy định chỉ thị về việc giãn cách xã hội dựa vào các phản ứng của cộng đồng trên mạng xã hội nói chung (cần thêm bộ lấy dữ liệu và tiền xử lý dữ liệu với các mạng xã hội ngoài phạm vi đề cương này) và mạng xã hội twitter nói riêng. ● <i>Thuật toán,</i> <p>Xây dựng hệ thống emotion detection tự động từ bộ dataset đã tạo ở Task 2 bằng phương pháp multi-channel CNN. Đầu tiên nhận dữ liệu đầu vào từ Task 1, tiến hành huấn luyện word embedding với kỹ thuật Word2Vec, đây là cách tiếp cận cung cấp nhiều tính đại diện của từ hơn là cách tiếp cận truyền thống word-base. Output của bước này là 1 vector 100 chiều. Sau đó sử dụng 1 block convolution để chiết xuất đặc</p>

	<p>trung, sau đó là 3 flatten layers để làm phẳng tensors. Cuối cùng là 1 dense layer để làm bộ phân lớp đầu ra (output classifier).</p> <ul style="list-style-type: none"> • <i>Bộ dữ liệu, etc</i> <p>Bộ dữ liệu nguyên thủy được crawl và nhận từ public APIs của Twitter với các tweets có các hashtag liên quan đến COVID-19 trong khoảng thời gian thực hiện các chính sách và chỉ thị về giãn cách xã hội để phòng chống dịch bệnh.</p> <p>Bộ dữ liệu được chia làm 2 phần với 80% dành cho mục đích training và 20% cho mục đích testing.</p>
Tài liệu tham khảo	<p>[1] Plutchik, R., A general psychoevolutionary theory of emotion, in Theories of emotion. 1980, Elsevier. p. 3–33</p> <p>[2] Kim, Y. Convolutional Neural Networks for Sentence Classification. Empirical Methods in Natural Language Processing (EMNLP). 2014</p> <p>[3] Aslam, F., et al., Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. Humanities and Social Sciences Communications, 2020. 7(1): p. 1–9</p> <p>[4] Li, Q., et al., Tracking and Analyzing Public Emotion Evolutions During COVID-19: A Case Study from the Event-Driven Perspective on Microblogs. IJERPH, 2020. 17(18): p. 6888</p> <p>[5] Khanpour, H. and C. Caragea. Fine-grained emotion detection in health-related online posts. Empirical Methods in Natural Language Processing. 2018</p>