I. MỞ ĐẦU

1. Tổng quan

Trong những năm gần đây việc phân tích và xử lý dữ liệu ngày càng trở nên phổ biến trong các lĩnh vực như kinh tế, công nghiệp,… Việc phân loại văn bản theo cảm xúc là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) và học máy (Machine Learning) như một bước tiến mới trong việc phát triển và đột phá về công nghệ.

2. Mục đích

Mục tiêu của lĩnh vực này là xác định và phân loại các cảm xúc hoặc thái độ thể hiện trong văn bản. Bài toán này thường liên quan đến việc gán nhãn cho các đoạn văn bản theo các cảm xúc khác nhau như tiêu cực, tích cực, hoặc trung lập.

Để xây dựng mô hình giải quyết bài toán nhận diện cảm xúc trong các văn bản tiếng Việt từ các ý kiến đánh giá, phản hồi,... cần đạt được những mục tiêu sau:

- Nhận diện tính tích cực – tiêu cực của văn bản.
- Xác định tính chủ quan – khách quan của văn bản.

Ngoài ra, mô hình giải quyết bài toán nhận diện cảm xúc trong văn bản tiếng Việt cần phải tối ưu về độ chính xác và hiệu suất thời gian thực hiện. Điều này nhằm giải quyết các vấn đề còn tồn tại trong việc nhận diện cảm xúc của khách hàng nói riêng và xử lý ngôn ngữ tự nhiên ở Việt Nam nói chung.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng quan trọng trong đề tài này là các câu đánh giá sản phẩm của người dùng. Các đánh giá phản hồi được khai thác trên Shopee đánh giá về các sản phẩm mà người dùng mua.

4. Phương pháp thực hiện

Trong đề tài này sử dụng các kiến trúc mạng học sâu là CNN – Convolutional Neural Networks và bi-LSTM - Bidirectional Long Short-Term Memory và các phương pháp tiền xử lý dữ liệu như loại bỏ dấu, tokenizer từ và văn bản

II. TẬP DỮ LIỆU

2.1. Tổng quan

Trong giới hạn môn học này, cụ thể hơn là bài tập này, chúng ta sẽ sử dụng tập dữ liệu về sắc thái các bình luận, đánh giá về các sản phẩm của khách hàng trên shopee.

Đặc tính	**Giá trị**
Số mẫu huấn luyện	22022
Số mẫu kiểm thử	9438
Ngôn ngữ	Tiếng Việt
Tổng số từ trong từ điển	9777
Độ dài đánh giá ngắn nhất	1 ký tự
Độ dài đánh giá tối đa	304 ký tự
Độ dài đánh giá trung bình	45 ký tự
Số lượng cảm xúc phân loại	3 (tiêu cực: NEG; trung bình: NEU; tích cực: POS)

2.2. Tiền xử lý

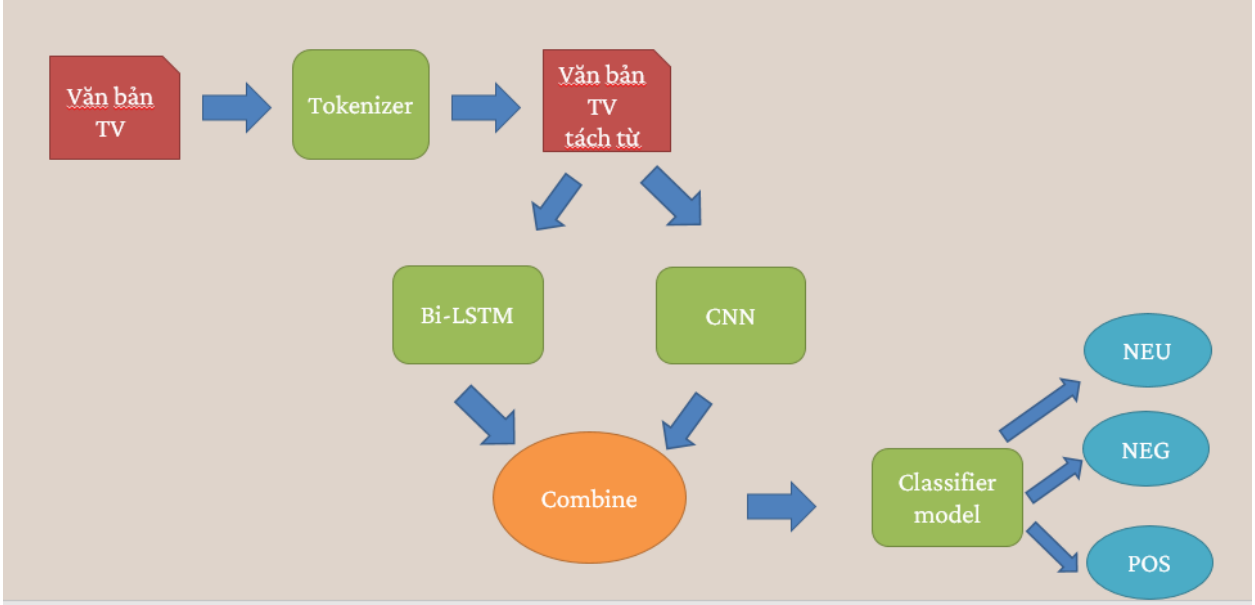
- Chuyển chữ hoa thành chữ thường
- Loại bỏ biểu tượng cảm xúc
- Loại bỏ URL
- Loại bỏ tên người và địa chỉ email
- Phân chia có dấu và không dấu

– Tokenize từng văn bản

2.3. Nhãn

– Chuyển đổi các nhãn từ dạng chuỗi sang dạng số

III. MÔ HÌNH HUẤN LUYỆN



Hình I. Mô hình huấn luyện

Theo sơ đồ trên dữ liệu đầu vào là các câu văn dạng văn bản có độ dài yêu cầu. Dữ liệu được chuyển thành các vectơ embedding. Một phần của dữ liệu cũng được đưa vào CNN để trích xuất đặc trưng cục bộ. Phần còn lại của dữ liệu được đưa vào mạng LSTM để xử lý tuần tự. Cả hai đặc trưng được kết hợp và đi qua các Classifier model để phân loại. Đầu ra là xác suất của các lớp nhãn (Negative, Neutral, Positive).

- Embedding: Sử dụng phương pháp nhúng từ cho các từ điển và one-hot coding cho nhãn
- CNN: Dùng CNN để lấy các đặc trưng của từng câu văn. Chỉ ra, phát hiện các cụm từ mang nghĩa tích cực/ tiêu cực/ trung tính
- Bi-LSTM: Học từ đầu câu đến cuối câu và ngược lại để ghi nhớ, dự đoán các từ tiếp theo và dự đoán mối quan hệ

- Kết hợp đặc trưng của CNN và sự học của Bi-LSTM thông qua các Classifier model để đưa ra kết quả cần đạt : Positive, Negative, Neutral
- **Sử dụng các thư viện NLP sẵn có:** Python có nhiều thư viện NLP như demoji, pyvi, scikit-learn có thể hỗ trợ các tác vụ xử lý ngôn ngữ tự nhiên, bao gồm phân loại văn bản theo cảm xúc.
- **Phát triển mô hình học máy/học sâu:** Sử dụng TensorFlow để phát triển mô hình phân loại văn bản theo cảm xúc tùy chỉnh.

IV. KẾT QUẢ ĐÁNH GIÁ

Link code kết quả: https://github.com/huuhoang129/Sentiment-Analysis

Sau khi áp dụng mô hình giải quyết bài toán phân loại văn bản theo cảm xúc gồm các bước: Tiền xử lý dữ liệu, vector hóa dữ liệu và phân loại cảm xúc bằng mô hình nhận diện cảm xúc sử dụng học sâu đã đạt được kết quả tương đối khả quan.

Sau khi huấn luyện và kiểm tra các dữ liệu ban đầu bằng phương pháp **Sentiment Analysis Vietnamese - SAV** đã cho kết quả khoảng từ **55% đến 75%**

Để làm được điều đó, cần phải hoàn thành được những việc như sau:

- Tìm hiểu về các đặc điểm của ngôn ngữ tiếng Việt, về xử lý ngôn ngữ tự nhiên và xử lý ngôn ngữ tiếng Việt. Tìm hiểu, phân tích và xây dựng thành công mô hình giải quyết bài toán phân lớp cảm xúc người dùng với định tính "Xác định tính tích cực – tiêu cực của văn bản".

- Tìm hiểu và áp dụng phương pháp vector hóa dữ liệu và CNN.

- Tìm hiểu các phương pháp tiền xử lý tiếng Việt nhằm cải thiện hiệu suất khi tiến hành huấn luyện.

- Tìm hiểu và áp dụng các phương pháp phân lớp và kết hợp với ba phương pháp xử lý văn bản tiếng Việt kể trên để chọn ra được phương pháp máy học tốt nhất cho phân lớp cảm xúc người dùng.

- Áp dụng kết hợp các phương pháp xử lý văn bản tiếng Việt và các thuật toán phân lớp để đánh giá trên bộ dữ liệu

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1]. Giáo trình "Xử lý ngôn ngữ tự nhiên", Đinh Điền, NXB Đại học Quốc gia – HCM, năm 2006.

[2]. Lê Si Lắc. "Nghiên cứu bài toán phân tích cảm xúc của người dùng". [2021]. Từ: https://fr.slideshare.net/slideshow/kha-lun-nghin-cu-bi-ton-phn-tch-cm-xc-ca-ngi-hng-9166421/250558592

[3]. Lê Minh Tứ. "Cài đặt mô hình phân loại cảm xúc tiếng Việt". [29 tháng 1 năm 2023]. Từ: https://viblo.asia/p/cai-dat-mo-hinh-phan-loai-cam-xuc-tieng-viet-018J2vdRJYK

[4]. SCV. "Phân tích cảm xúc trong Tiếng Việt". Từ: https://streetcodevn.com/blog/sav

[5]. Word embeddings. Từ: https://www.tensorflow.org/text/guide/word_embeddings

[6]. Huy Lê Vũ Minh. "Phân loại cảm xúc văn bản". Từ: https://www.youtube.com/watch?v=rckCSs0GiKA

[7] ProtonX. "Xây dựng mô hình phân loại cảm xúc và trích xuất embedding của từ (Train Sentiment Classification)". Từ: https://www.youtube.com/watch?v=JIafLwlGzBA&t=501s