ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

NGUYỄN THỊ THU HÒA

NGHIÊN CỨU MỘT SỐ KỸ THUẬT NLP VÀ ỨNG DỤNG PHÂN LOẠI VĂN BẢN TIẾNG VIỆT

LUẬN VĂN THẠC SĨ
Chuyên ngành: Khoa học máy tính
Mã số: 8480101

Người hướng dẫn: TS Trần Quang Quý

Thái Nguyên - 2023

LỜI CẢM ƠN

Em xin gửi lời cảm ơn sâu sắc tới thầy giáo hướng dẫn TS.Trần Quang Quý. Thầy đã giao đề tài và tận tình hướng dẫn em trong quá trình hoàn thành đề tài này.

Em xin gửi lời cám ơn của mình tới các thầy cô giáo trong trường Đại học Công nghệ thông tin và Truyền Thông - ĐHTN, các thầy cô đã giảng dạy giúp đỡ em trong quá trình học tập tại trường.

Tôi xin chân thành cảm ơn các đồng nghiệp ở cơ quan, trường Cao đẳng Lào Cai. Cảm ơn gia đình bè bạn đã giúp đỡ tôi hoàn thành luận văn này.

Thái Nguyên, ngày 10 tháng 8 năm 2023

Sinh Viên

Nguyễn Thị Thu Hòa

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do chính tôi thực hiện, dưới sự hướng dẫn của TS.Trần Quang Quý. Các kết quả lý thuyết được trình bày trong luận văn là sự tổng hợp từ các kết quả đã được công bố và có trích dẫn đầy đủ, kết quả của chương trình thực nghiệm trong luận văn này được tác giả thực hiện là hoàn toàn trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm.

Thái Nguyên, ngày 10 tháng 8 năm 2023

Học viên

Nguyễn Thị Thu Hòa

Mục lục

Danh sách bảng

Danh sách hình vẽ

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

NLP	Xử lý ngôn ngữ tự nhiên
Word Representation	Biểu diễn từ.
Tokenize	Tách từ.
$\|\mathbf{x}\|_2$	Chuẩn 2 của véc tơ, chuẩn Euclide
$\mathbb{R}^n$	Không gian véc tơ thực n chiều.
Sklearn	Thư viện học máy cơ bản trong python.
Word embedding	Nhúng từ.
SVD	Phân tích ma trận bằng phương pháp suy biến.
PCA	Một phương pháp giảm chiều, phân tích thành phần chính.
Word2vec	Một phương pháp đưa từ thành véc tơ.
Layers	Các lớp, tầng trong mạng nơ ron.
Skip-Grams	Một phương pháp nhúng từ.
$P(x\|y)$	Xác xuất xuất hiện x với điều kiện y.
Layer	Tầng, lớp mạng.
CBOW	Một mô hình nhúng từ.
TensorFlow	Một framework hỗ trợ thực hành deeplearning bằng python.
epochs	Số lần huấn luyện mô hình.
t-SNE	Một thuật toán giảm chiều dùng để biểu diễn từ.
gensim	Một thư viện xử lý ngôn ngữ tự nhiên.
Bert	Mô hình biểu diễn từ hai chiều ứng dụng Transformer
Attention	Cơ chế chú ý
fine tuning	Tinh chỉnh mô hình (tham số)
MLM	Mặt nạ LM - tác vụ tinh chỉnh biểu diễn từ
Fully connect	Kết nối đầy đủ
SOTA	State-of-the-art Kết quả tốt nhất
Roberta	Một kiến trúc, thuật toán của facebook trên pytorch.
BPE	Mã hóa Byte Pair Encoding

MỞ ĐẦU

Xử lý ngôn ngữ tự nhiên (NLP) đại diện cho một phân nhánh quan trọng trong lĩnh vực Trí tuệ nhân tạo, đặc biệt chú trọng vào việc khám phá sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người, bất kể đó là qua hình thức tiếng nói (speech) hay văn bản (text) [1], [2]. Mục tiêu vượt qua trong lĩnh vực này là giúp máy tính thấu hiểu và thực hiện các nhiệm vụ liên quan đến ngôn ngữ của con người một cách hiệu quả, bao gồm việc tương tác giữa con người và máy, cải thiện chất lượng giao tiếp giữa con người và con người, cũng như tối ưu hóa quá trình xử lý cả văn bản và lời nói.

Xử lý ngôn ngữ tự nhiên có nguồn gốc từ những năm 1940 của thế kỷ 20, trải qua một hành trình phát triển đa dạng với một loạt phương pháp và mô hình xử lý khác nhau. Các giai đoạn quan trọng bao gồm việc áp dụng các phương pháp ô-tô-mát và mô hình xác suất vào những năm 1950, việc khai thác ký hiệu và các phương pháp ngẫu nhiên trong những năm 1970, tiến bộ sử dụng học máy truyền thống vào đầu thế kỷ 21, và đặc biệt là sự bùng nổ của học sâu trong thập kỷ gần đây [3], [4].

Xử lý ngôn ngữ tự nhiên có khả năng được phân chia thành hai hướng phát triển quan trọng, không hoàn toàn độc lập, gồm xử lý tiếng nói (speech processing) và xử lý văn bản (text processing). Xử lý tiếng nói tập trung vào việc nghiên cứu và phát triển các thuật toán cũng như chương trình máy tính nhằm xử lý ngôn ngữ con người ở dạng tiếng nói (dữ liệu âm thanh). Trong lĩnh vực này, nhận dạng tiếng nói và tổng hợp tiếng nói đứng ra như những ứng dụng quan trọng. Nhận dạng tiếng nói liên quan đến việc chuyển đổi ngôn ngữ từ dạng tiếng nói sang dạng văn bản, trong khi tổng hợp tiếng nói thực hiện chuyển ngôn ngữ từ dạng văn bản thành tiếng nói.

Xử lý văn bản tập trung vào việc phân tích dữ liệu văn bản. Các ứng dụng quan trọng trong lĩnh vực này bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động và kiểm tra chính tả tự động. Thậm chí, xử lý văn bản có thể tiếp tục được chia thành hai phân nhánh nhỏ hơn, bao gồm hiểu văn bản và sinh văn bản. Hiểu văn bản liên quan đến các nhiệm vụ phân tích văn bản, trong khi sinh văn bản liên quan đến việc tạo ra văn bản mới, như trong các ứng dụng dịch máy hoặc tóm tắt văn bản tự động [4].

Bài toán phân loại văn bản là một trong những bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Quy trình cơ bản để xây dựng mô hình phân loại văn bản sử dụng các mô hình học máy, học sâu là [4]:

- Chuẩn bị dữ liệu: Thu thập và chuẩn bị dữ liệu cho bài toán phân loại văn bản. Dữ liệu này gồm các văn bản đã được gán nhãn (label) cho từng loại phân loại. Ví dụ: nếu muốn phân loại email là spam hay không, dữ liệu sẽ bao gồm các email đã được gán nhãn là "spam" hoặc "không phải spam".

- Tiền xử lý dữ liệu: Dữ liệu văn bản thường cần được tiền xử lý để loại bỏ các ký tự đặc biệt, chuyển đổi về dạng chuẩn (ví dụ: chuyển đổi thành chữ thường), loại bỏ từ dừng (stop words) và thực hiện các quy trình như tokenization, stemming hoặc lemmatization.

- Trích xuất đặc trưng: Văn bản cần được biểu diễn thành các đặc trưng số học để máy học có thể xử lý. Một số phương pháp phổ biến để trích xuất đặc trưng từ văn bản là sử dụng bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, hoặc BERT (Bidirectional Encoder Representations from Transformers).

- Xây dựng mô hình: Chọn một thuật toán phù hợp để huấn luyện mô hình phân loại. Các thuật toán phổ biến dùng trong phân loại văn bản bao gồm Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Random Forest, hoặc các mô hình học sâu như Recurrent Neural Networks (RNNs) và Convolutional Neural Networks (CNNs).

- Huấn luyện và đánh giá mô hình: Sử dụng dữ liệu đã được gán nhãn, huấn luyện mô hình trên tập huấn luyện và đánh giá hiệu suất của mô hình trên tập kiểm tra. Các phép đo đánh giá thông thường bao gồm độ chính xác (accuracy), độ phủ (recall), độ chính xác trung bình trọng số (weighted precision), và F1-score.

- Tinh chỉnh mô hình

Bài toán phân loại văn bản tiếng Việt có một số điểm khác biệt so với phân loại văn bản nói chung do đặc thù ngôn ngữ và ngữ cảnh văn hóa. Một số điểm khác biệt quan trọng [9]:

- Ngôn ngữ và Từ vựng: Ngôn ngữ tiếng Việt có các đặc điểm riêng, bao gồm cấu trúc ngữ pháp, từ vựng và cách ngữ âm. Điều này tạo ra một số thách thức trong việc biểu diễn từ vựng, xử lý ngữ pháp và thậm chí việc phát hiện từ ngữ mới hay viết sai.

- Từ viết tắt và biểu ngữ cụ thể: Ngôn ngữ tiếng Việt thường sử dụng rất nhiều từ viết tắt, biểu ngữ cụ thể và ngôn ngữ thông tin (ví dụ: "e", "mình", "mk" thay cho "tôi", "mình"). Điều này có thể làm cho việc biểu diễn ngôn ngữ và phân loại trở nên phức tạp hơn.