Multi-extended MinBERT for sentiment classification task with contrastive learning

Huy Tran Vu Duc
University of Engineering and Technology, VNU
Hanoi, Vietnam
22021111@vnu.edu.vn

Duong Tran Anh
University of Engineering and Technology, VNU
Hanoi, Vietnam
22028334@vnu.edu.vn

Minh Le Ba Quang
University of Engineering and Technology, VNU
Hanoi, Vietnam
22021222@vnu.edu.vn

***Tóm tắt nội dung*—Hiện nay, để giải quyết vấn đề phân tích cảm xúc (sentiment analysis), nhiều mô hình NLP dựa trên học chuyển giao đã được đề xuất. Tuy nhiên, hầu hết các mô hình này đều chỉ tập trung vào phân loại nhị phân. Bên cạnh đó, việc tinh chỉnh một mô hình lớn như BERT [1] cũng thường đối mặt với những khó khăn như tình trạng quá khớp (overfitting) hay tình trạng nhúng từ dị hướng (anisotropic word embeddings), làm giảm khả năng phân biệt các câu của mô hình dựa trên các đặc điểm như cảm xúc. Trong dự án này, nhóm khám phá sự kết hợp các kỹ thuật để nâng cao hiệu suất của mô hình BERT đã đào tạo trước (pre-trained) trên tác vụ phân tích cảm xúc chi tiết (fine-grained sentiment analysis). Trong phần đầu của dự án, nhóm đã áp dụng các kỹ thuật học tương phản (SimCSE của Gao và cộng sự [2]) giám sát và không giám sát vào BERT để giảm tính dị hướng của nhúng câu (sentence embedding). Ở phần thứ hai, nhóm đã đưa chính quy hóa (regularization) vào hàm mất mát tinh chỉnh (fine-tuning loss) và cập nhật gradient tham số (SMART của Jiang và cộng sự [3]) nhằm cải thiện tình trạng quá khớp. Sau đó, nhóm thực hiện so sánh hiệu quả của mô hình BERT baseline với những mô hình ở hai phần của dự án. Kết quả cho thấy mô hình BERT với cải tiến sử dụng kỹ thuật của SMART có hiệu suất tốt nhất trong các mô hình.**

***Index Terms*—Phân tích cảm xúc chi tiết, Học chuyển giao, Học tương phản, BERT**

I. INTRODUCTION

Phân tích cảm xúc là tác vụ phân loại một đoạn văn bản vào một trong các lớp cảm xúc được xác định trước. Trong phân tích cảm xúc chi tiết, có năm lớp là: rất tiêu cực, tiêu cực, trung tính, tích cực và rất tích cực. Kể từ khi Vaswani và cộng sự (2017) [4] giới thiệu kiến trúc Transformer, học chuyển giao trên các mô hình như BERT (Devlin và cộng sự, 2019) [1], đã khẳng định được vị thế là nền tảng để đạt được kết quả tiên tiến (SOTA) trên nhiều tác vụ NLP. Điểm chuẩn Stanford Sentiment Treebank (Socher và cộng sự, 2013) [5] cho thấy sự cải thiện hiệu suất đáng kể thông qua việc điều chỉnh theo từng tác vụ cụ thể của các mô hình này. Khi phát hành, BERT đã đạt được kết quả tiên tiến trên 11 tác vụ xử lý ngôn ngữ tự nhiên, đẩy điểm GLUE [6] lên 80,5% (cải thiện 7,7%) (Devlin và cộng sự, 2019) [1]. Tuy nhiên, BERT vẫn còn những lĩnh vực cần cải thiện như phân tích cảm xúc chi tiết. Liệu hiệu suất của tác vụ này có thể được cải thiện bằng việc khắc phục vấn đề trong học chuyển giao nêu sau đây?

Thứ nhất, các mô hình lớn được đào tạo trước này không phải lúc nào cũng tạo ra các nhúng tối ưu cho mọi tác vụ. Các công trình trước đây chỉ ra rằng các nhúng câu từ các mô hình ngôn ngữ được đào tạo trước có thể bị giới hạn bởi không gian nhúng từ dị hướng đã học được. Do đó, đôi khi các mô hình này gặp khó khăn khi khái quát hóa cho các tác vụ hạ nguồn. Kể từ khi các mô hình ngôn ngữ lớn xuất hiện, đã có rất nhiều công trình xung quanh cách thức đào tạo trước và tinh chỉnh các nhúng này để tạo ra kết quả tốt hơn. Trong SimCSE: Simple Contrastive Learning of Sentence Embeddings [2] đã triển khai học tương phản cho BERT để đạt được điểm cao hơn về điểm tương đồng ngữ nghĩa. Lấy cảm hứng từ cách tiếp cận này, trước tiên nhóm muốn xem tác dụng của việc triển khai học tương phản trên BERT. Thứ hai, rất khó để thực hiện đúng quá trình khuyến khích một mô hình học một tác vụ mới hiệu quả mà không "quên" kiến thức ngôn ngữ khó có được trước đó. Ở đây, Jiang và cộng sự (2020) [3] đề xuất kỹ thuật điều chỉnh SMART - một cơ chế chống lại vấn đề quá khớp liên quan đến học chuyển giao.

Trong dự án này, mục tiêu của nhóm là triển khai một mô hình BERT có thể hoạt động tốt trên tác vụ phân tích cảm xúc chi tiết. Cụ thể hơn, nhóm cố gắng kết hợp ý tưởng học tương phản và chính quy hóa ở trên vào kiến trúc với hy vọng cải thiện hiệu suất ban đầu. Để thực hiện như vậy, đầu tiên nhóm thực hiện tinh chỉnh riêng biệt mô hình BERT với phiên bản học giám sát và không giám sát của simCSE. Sau đó, nhóm triển khai SMART nhằm giải quyết các vấn đề quá khớp trong quá trình tinh chỉnh. Cuối cùng, nhóm thực hiện tìm kiếm các siêu tham số của mô hình bao gồm tốc độ học, xác suất dropout, siêu tham số trong simCSE và SMART.

Phần còn lại của báo cáo gồm 7 phần. Phần II là một bản tóm tắt về các công trình liên quan. Trong phần III, nhóm giải thích chi tiết về kiến trúc mô hình của nhóm. Tiếp theo, ở phần IV sẽ mô tả các tập dữ liệu mà nhóm sử dụng và cách nhóm thực hiện tiền xử lý dữ liệu. Sau đó, nhóm trình bày các thí nghiệm mà nhóm thực hiện trong phần V và kết quả của các thí nghiệm này trong phần VI. Cuối cùng, nhóm đưa ra phân tích về kết quả thí nghiệm trong phần VII và nhận xét kết luận cho công trình trong phần VIII.

II. RELATED WORKS

Phân tích cảm xúc là một trong những nhiệm vụ phổ biến nhất trong NLP đã có rất nhiều nghiên cứu và tiến bộ trong việc giải quyết nhiệm vụ này. Trước tiên nhóm đi qua một số phương pháp học sâu được áp dụng cho phân tích cảm xúc. Tai và cộng sự (2015) [7] đã áp dụng nhiều dạng LSTM khác nhau và Kim (2014) [8] đã áp dụng CNN để phân tích cảm xúc. Các phương pháp này đều không theo ngữ cảnh, tức là chúng tạo ra nhúng từ duy nhất cho mỗi từ trong vốn từ vựng. Nghiên cứu mô hình ngôn ngữ gần đây đã cố gắng đào tạo nhúng theo ngữ cảnh. Peters và cộng sự (2018) [9] đã trích xuất ngữ cảnh dựa trên LSTM từ trái sang phải và từ phải sang trái. Devlin và cộng sự (2019) [1] đã đề xuất BERT (Bidirectional Encoder Representations from Transformers), một kiến trúc Transformer dựa trên sự chú ý (attention) [4], để đào tạo các biểu diễn hai chiều từ các văn bản không có nhãn. Kiến trúc của họ không chỉ thu được kết quả tiên tiến nhất trên nhiều tác vụ NLP mà còn cho phép mức độ song song cao vì nó không dựa trên các kết nối tuần tự hoặc tuần hoàn.

Bên cạnh đó, sự thành công của các kỹ thuật NLP phụ thuộc vào lượng lớn dữ liệu được gắn nhãn. Hầu hết các phương pháp hiện nay đều tập trung vào phân loại nhị phân, rất có thể là do có các tập dữ liệu công khai lớn dành cho nó như tập dữ liệu đánh giá phim IMDb. Để giải quyết hạn chế về lượng dữ liệu được gắn nhãn chi tiết trong phân tích cảm xúc, các nhà nghiên cứu đề xuất phương pháp học chuyển giao. Học chuyển giao xem xét kịch bản, trong đó chúng ta có dữ liệu được gắn nhãn hạn chế từ miền đích cho một tác vụ nhất định, nhưng chúng ta có các tác vụ có liên quan với lượng lớn dữ liệu từ các miền khác nhau. Mục tiêu là chuyển kiến thức từ miền có nhiều tài nguyên sang miền đích có ít tài nguyên. Trong NLP, hầu hết các phương pháp học chuyển giao đều chọn đào tạo trước một mô hình ngôn ngữ lớn, ví dụ: ELMo (Peters và cộng sự, 2018) [9], GPT (Radford và cộng sự, 2019) [10] và BERT (Devlin và cộng sự, 2019) [1]. Mô hình ngôn ngữ như vậy có thể nắm bắt thông tin ngữ nghĩa và cú pháp chung có thể được sử dụng thêm trong các tác vụ NLP hạ lưu. Mô hình ngôn ngữ đặc biệt hấp dẫn vì nó có thể được đào tạo theo cách hoàn toàn không giám sát với lượng lớn dữ liệu không có nhãn, hiện nay rất rẻ để lấy từ internet. Ngữ liệu văn bản đa miền cực lớn thu được cho phép chúng ta đào tạo các mô hình ngôn ngữ khổng lồ.

Gần đây, nghiên cứu của Ethayarajh (2019) [11] cho thấy trong các biểu diễn ngữ cảnh của mô hình được đào tạo trước gặp vấn đề dị hướng. Thay vì được phân bổ đều, các nhúng chiếm một hình nón hẹp trong không gian vectơ. Wang và Isola (2020) [12] cho thấy mối liên quan giữa tính dị hướng với tính căn chỉnh (alignment) và tính đồng nhất (uniformity). Trong một tập nhúng được căn chỉnh tốt, các cặp có liên quan về mặt ngữ nghĩa chiếm các điểm gần nhau trong không gian vectơ. Điều này cho phép mô hình vẽ các mối quan hệ. Trong một tập nhúng đồng nhất, toàn bộ tập nhúng sẽ được phân bổ đồng đều trên toàn bộ phạm vi đơn vị. Điều này tối đa hóa lượng thông tin có thể được mã hóa trong các nhúng. Theo trực giác, việc tối ưu hóa mục tiêu học tập tương phản có thể cải thiện tính đồng nhất (hoặc làm giảm vấn đề dị hướng). Học tương phản nhằm mục đích học cách biểu diễn hiệu quả bằng cách kéo những người hàng xóm gần về mặt ngữ nghĩa lại với nhau và đẩy những người không phải là hàng xóm ra xa (Hadsell và cộng sự, 2006) [13]. Một khó khăn phát sinh khi triển khai học tương phản trong mô hình NLP là tính rời rạc khiến việc tạo ra các cặp tích cực và tiêu cực khó hơn nhiều. Mặc dù chúng ta có thể ngẫu nhiên bỏ qua một từ hoặc thay thế nó bằng từ đồng nghĩa, nhưng việc làm như vậy không phải lúc nào cũng đảm bảo một câu mới mạch lạc hoặc sẽ có ý nghĩa tương tự về mặt ngữ cảnh. Gao và cộng sự (2021) [2] đã đưa ra một khuôn khổ mới sử dụng học tập tương phản dẫn đến các nhúng đồng nhất hơn và căn chỉnh tốt hơn.

Ngoài ra, do dữ liệu hạn chế từ tác vụ/miền mục tiêu và độ phức tạp cao của mô hình được đào tạo trước, việc tinh chỉnh mạnh thường khiến mô hình được điều chỉnh quá mức so với dữ liệu đào tạo của tác vụ/miền mục tiêu và do đó không khái quát hóa tốt đối với dữ liệu chưa biết. Để giảm thiểu vấn đề này, các phương pháp tinh chỉnh thường dựa vào phương pháp tìm kiếm điều chỉnh siêu tham số. Ví dụ, Howard và Ruder (2018) [14] lập lịch tốc độ học theo phương pháp heuristic và dần dần giải phóng các lớp của mô hình ngôn ngữ để cải thiện hiệu suất tinh chỉnh, Peters và cộng sự (2019) [15] chỉ điều chỉnh một số lớp nhất định và đóng băng các lớp khác, Stickland và Murray (2019) [16] đề xuất thêm các lớp bổ sung vào mô hình được đào tạo trước và tinh chỉnh cả hai lớp hoặc chỉ các lớp bổ sung. Tuy nhiên, các phương pháp này đòi hỏi những nỗ lực tinh chỉnh đáng kể. Jiang và cộng sự (2019) [3] đề xuất một kỹ thuật chính quy hóa (SMART) hoạt động bằng cách giới thiệu một hàm mất mát đối nghịch và tạo vùng tin cậy để cập nhật tham số. Cách tiếp cận này không yêu cầu điều chỉnh siêu tham số đáng kể.

III. MODEL

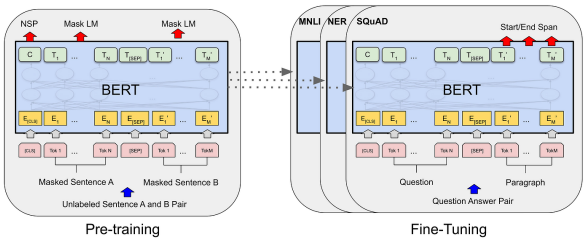
A. Base model

BERT (*Bidirectional Encoder Representations from Transformers*) được xây dựng dưới dạng một mô hình **Transformer** hai chiều, với kiến trúc bao gồm nhiều lớp **Encoder**. Điểm đặc trưng nổi bật của BERT nằm ở khả năng học biểu diễn ngữ cảnh hai chiều, cho phép mô hình nắm bắt thông tin từ cả bên trái và bên phải của một từ trong ngữ cảnh câu.

Quy trình huấn luyện của BERT bao gồm hai giai đoạn chính: **Pre-training** và **Fine-tuning**. Trong giai đoạn **Pre-training**, mô hình được huấn luyện trên một lượng lớn dữ liệu không gán nhãn bằng cách sử dụng hai nhiệm vụ chính là *Masked Language Modeling* (MLM) và *Next Sentence Prediction* (NSP). Giai đoạn này nhằm mục đích học các biểu diễn ngữ nghĩa tổng quát từ dữ liệu. Sau đó, trong giai đoạn **Fine-tuning**,

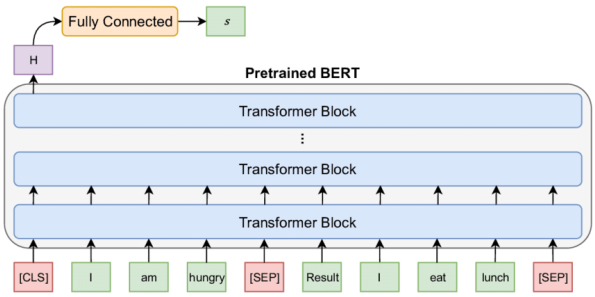
mô hình được điều chỉnh lại trên các tập dữ liệu đặc thù cho từng tác vụ cụ thể, chẳng hạn như phân loại văn bản, nhận diện thực thể hoặc trả lời câu hỏi.

Hình 1 minh họa chi tiết hai giai đoạn huấn luyện của mô hình BERT, từ giai đoạn tiền huấn luyện đến tinh chỉnh.



Hình 1. Hai bước huấn luyện mô hình BERT

Nhóm sử dụng **BertForSequenceClassification** có sẵn trong thư viện **Hugging Face Transformers**, đây là một lớp được xây dựng dựa trên mô hình **BERT** (Nhóm sử dụng với phiên bản $BERT_{BASE}$). Bằng việc thêm một lớp **fully connected layer (classification head)** trên đầu mô hình BERT và sử dụng embedding của token [CLS] đại diện cho toàn bộ chuỗi đầu vào, kết quả thu được là vector có kích thước (batch_size, num_labels) để dự đoán nhãn. Mô hình **BertForSequenceClassification** được thể hiện ở hình 2



Hình 2. BertForSequenceClassification

B. Our extensions

1) SimCSE: Là một phương pháp học tương phản nhằm cải thiện chất lượng của các câu embedding, giúp các câu được embedding trở nên tách biệt rõ ràng hơn trong không gian vector biểu diễn. Phương pháp này sử dụng cơ chế học đối chiếu (contrastive learning) để tối ưu hóa khoảng cách giữa các embedding của các câu tương tự, đồng thời tăng cường sự phân biệt giữa các câu không tương tự, từ đó nâng cao khả năng phân loại và truy vấn ngữ nghĩa trong không gian vector. Hình 4 mô tả hai phương pháp học tương phản của nhóm.

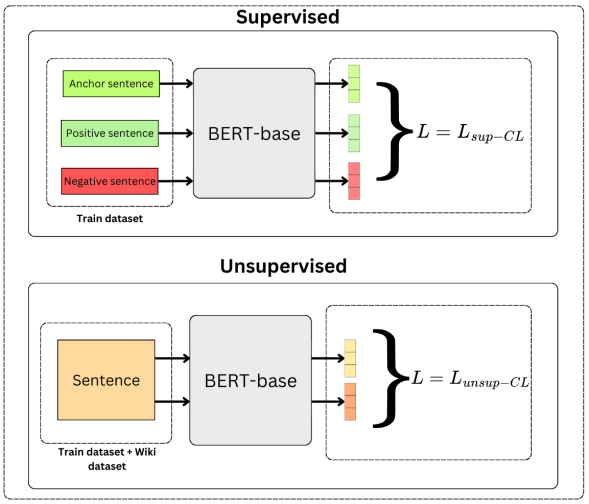
Supervised learning: Là phương pháp học tương phản có giám sát. Với ý tưởng của nhóm là tạo một dataset theo format của NLI dataset trong paper gốc của SimCSE với mỗi sample gồm 3 câu (sent0, sent1, hard-neg) lần lượt là 2 câu cùng level sentiment và 1 câu với level sentiment liền kề. Nhóm huấn luyện với hàm mục tiêu :

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}(e^{sim(\mathbf{h}_i,\mathbf{h}_j^+)/\tau} + e^{sim(\mathbf{h}_i,\mathbf{h}_j^-)/\tau})}$$

Với $sim(\mathbf{h}_1, \mathbf{h}_2) = \frac{\mathbf{h}_1^T \mathbf{h}_2}{||\mathbf{h}_1|| \cdot ||\mathbf{h}_2||}$ và (sent0, sent1, hard-neg) có embedding vector là ($\mathbf{h}$, $\mathbf{h}^+$, $\mathbf{h}^-$)

Unsupervised learning: Là phương pháp học tương phản không giám sát. Với ý tưởng theo bài báo gốc, đưa một câu qua mô hình với dropout mask khác nhau để lấy được 2 embedding vectors cùng biểu diễn 1 câu đầu vào (thường dùng cùng tỷ lệ dropout). Nhóm huấn luyện với hàm mục tiêu :

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i^{z_i},\mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{h}_i^{z_i},\mathbf{h}_j^{z_j'})/\tau}}$$

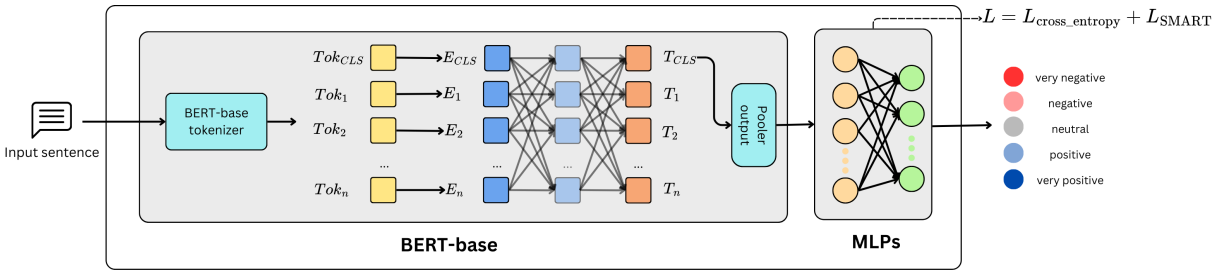


Hình 4. Phương pháp học tương phản của nhóm

2) Multilayer Perceptrons: Sau khi mô hình được huấn luyện để cải thiện các vector biểu diễn (embeddings), nhóm đưa các vector này qua một mạng Multilayer Perceptron (MLP), bao gồm các lớp fully connected, Batch Normalization, Dropout và hàm activation ReLU. Quá trình này nhằm mục đích tăng cường khả năng học biểu diễn phi tuyến và giảm thiểu overfitting, đồng thời cải thiện hiệu suất tổng quát của mô hình trong việc học các đặc trưng ngữ nghĩa từ dữ liệu.

3) SMART: Sự khan hiếm dữ liệu huấn luyện có thể dẫn đến các cập nhật mô hình quá mạnh trong quá trình tinh chỉnh, dẫn đến hiện tượng overfitting. SMART cung cấp hiệu ứng điều chuẩn mạnh mẽ đối với các cập nhật mô hình thông qua hai chiến lược chính:

Smoothness-Inducing Adversarial Regularization: Nhóm áp dụng kỹ thuật regularization để cải thiện độ ổn định của các vector biểu diễn trong mô



Hình 3. Mô hình tổng quát

hình học sâu, thông qua việc thêm noise vào dữ liệu đầu vào trong quá trình huấn luyện. Điều này giúp giảm overfitting, cải thiện khả năng tổng quát của mô hình và đảm bảo các vector biểu diễn phản ánh chính xác mối quan hệ ngữ nghĩa giữa các câu. Kỹ thuật này không chỉ giúp mô hình ổn định hơn mà còn nâng cao khả năng học các đặc trưng ngữ nghĩa bền vững.
Nhóm sẽ đi giải quyết bài toán tối ưu:

$$\begin{aligned}\min_{\theta} \mathcal{F}(\theta) &= \mathcal{L}(\theta) + \lambda_s \mathcal{R}_s(\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i,\theta), y_i)) + \\ \lambda_s \frac{1}{n}\sum_{i=1}^{n} &\max_{||\tilde{x_i} - x_i||_p \le \epsilon} \ell_s(f(\tilde{x_i},\theta), f(x_i,\theta)) \end{aligned} \quad (1)$$

trong đó : $\mathcal{L}(\theta)$ là hàm mất mát gốc, $\mathcal{R}_s(\theta)$ là hàm mất mát SMART, ℓ_s là hàm Phân kỳ Kullback–Leibler đối xứng.

Bregman Proximal Point Optimization: Là một phương pháp tối ưu tiên tiến nhằm kiểm soát quá trình cập nhật gradient, *hạn chế việc cập nhật quá mức* tại mỗi vòng lặp của thuật toán. Kỹ thuật này được áp dụng để giải quyết bài toán tối ưu hóa hàm mất mát (1), cung cấp một cơ chế điều chỉnh tinh vi giúp ổn định quá trình học và cải thiện hội tụ của mô hình.

Nhóm sử dụng mô hình pre-trained làm khởi tạo, ký hiệu là $f(.,\theta_0)$. Ở lần lặp thứ (t+1) thì sẽ xác định:

$$\begin{aligned}\theta_{t+1} &= arg\min_{\theta} \mathcal{F}(\theta) + \mu \mathcal{D}_{Breg}(\theta, \tilde{\theta_t}) \\ &= arg\min_{\theta} \mathcal{F}_\theta + \mu \frac{1}{n}\sum_{i=1}^{n} \ell_s(f(x_i,\theta), f(x_i,\tilde{\theta_t}))\end{aligned}$$

Với $\beta \in [0,1)$ và $\tilde{\theta_t} = (1-\beta)\theta_t + \beta\tilde{\theta}_{t-1}$

Hình 3 minh họa mô hình phân loại cảm xúc dựa trên BERT. Câu đầu vào được chuyển thành các token thông qua BERT-base tokenizer và sau đó được mã hóa bởi mô hình BERT-base để tạo ra các vector biểu diễn, trong đó token [CLS] được sử dụng làm biểu diễn tổng quát của toàn câu. Vector biểu diễn của [CLS] tiếp tục được xử lý qua tầng "pooler" để tạo đầu vào cho mạng MLPs (Multi-Layer Perceptrons), thực hiện phân loại cảm xúc thành 5 nhãn: very negative, negative, neutral, positive, very positive. Quá trình huấn luyện sử dụng hàm mất mát tổng hợp bao gồm Cross Entropy loss cho phân loại và SMART loss nhằm cải thiện tính ổn định và khả năng tổng quát hóa của mô hình.

IV. DATASETS

A. Tập dữ liệu

Mô hình BERT nhóm sử dụng được đào tạo trước bằng hai tác vụ không giám sát trên các bài viết Wikipedia: MLM (Masked language modeling) và NSP (Next sentence prediction). Để thực hiện tinh chỉnh trên tác vụ phân tích tình cảm chi tiết, nhóm sử dụng các tập dữ liệu Stanford Sentiment Treebank (SST) [5], Amazon Product Reviews [1] và wiki1m_for_ simcse [2].

wiki1m_for_simcse là tập dữ liệu được sử dụng cho phiên bản học không giám sát trong bài báo simCSE. Tập dữ liệu gồm 10^6 câu được lấy mẫu ngẫu nhiên từ English Wikipedia. Với tập dữ liệu này, nhóm đã trích ra tập dữ liệu con sample-wiki gồm 443573 câu có độ dài lớn hơn 256 ký tự để sử dụng cho phiên bản simCSE không giám sát của nhóm.

Stanford Sentiment Treebank (SST) gồm các câu đơn lấy từ các bài đánh giá phim trên Rotten Tomatoes. Bộ dữ liệu được phân tích cú pháp bằng trình phân tích cú pháp Stanford và gồm 215.154 cụm từ duy nhất từ các cây phân tích cú pháp đó. Hơn nữa, mỗi nốt được dán 1 trong 5 nhãn là {0, 1, 2, 3, 4} bởi ít nhất ba người.

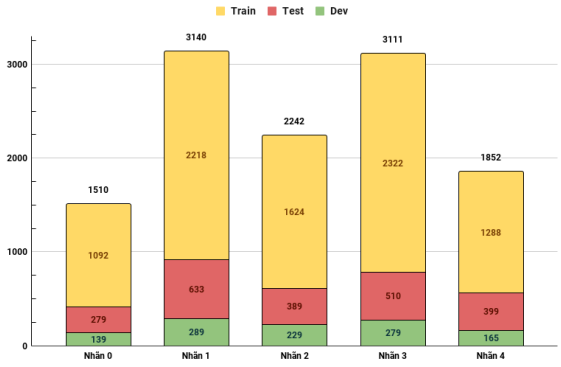
Amazon Product Reviews gồm 21000 câu đánh giá sản phẩm với 5 nhãn là {1, 2, 3, 4, 5}. Tập dữ liệu được chia thành 5 thư mục tương ứng với 5 nhãn. Mỗi thư mục bao gồm 4200 tệp văn bản chứa một đánh giá riêng lẻ. Tuy nhiên, nhóm nhận thấy tập dữ liệu Amazon Product Reviews có rất nhiều câu bị lặp. Vì vậy nhóm đã thực hiện xóa các câu bị lặp và sửa lại cấu trúc thư mục thành tập dữ liệu mới là amazon-reviews.

Thông tin chi tiết về việc chia dữ liệu của các tập dữ liệu ở bảng I, hình 5 và hình 6

[1]https://www.kaggle.com/datasets/olegoleshchuk/productreviews
[2]https://huggingface.co/datasets/princeton-nlp/datasets-for-simcse/resolve/main/wiki1m_for_simcse.txt

	SST			amazon-reviews		
	Dev	Test	Train	Dev	Test	Train
Nhãn 0	139	279	1092	166	331	1158
Nhãn 1	289	633	2218	125	250	871
Nhãn 2	229	389	1624	99	198	692
Nhãn 3	279	510	2322	200	400	1400
Nhãn 4	165	399	1288	123	245	854
Tổng số câu	1101	2210	8544	713	1424	4975

Bảng I
THÔNG KÊ DỮ LIỆU STANFORD SENTIMENT TREEBANK VÀ FILTERED-AMAZON



Hình 5. Phân bố dữ liệu Stanford Sentiment Treebank

Hình 6. Phân bố dữ liệu amazon-reviews

B. Tiền xử lý

Nhóm thực hiện các bước tiền xử lý sau đây trên các tập dữ liệu trước khi đưa chúng vào mô hình của nhóm.

1) **Chuẩn hóa (Canonicalization):** Đầu tiên, nhóm xóa tất cả các chữ số, ký hiệu dấu câu, dấu trọng âm, khoảng trắng và chuyển đổi thành chữ thường.
2) **Mã hóa (Tokenization):** Sau đó, nhóm mã hóa văn bản bằng Bert_base tokenizer. Các từ được chia nhỏ thành tiền tố, gốc và hậu tố từ đó giúp xử lý các từ chưa thấy tốt hơn. Ví dụ: playing → play + ##ing.
3) **Thêm token đặc biệt:** Cuối cùng, nhóm thêm token [CLS] và [SEP] ở các vị trí thích hợp.

V. EXPERIMENTS

A. Evaluation method

Nhóm thực hiện huấn luyện mô và đánh giá hình đề xuất trên các bộ dữ liệu SST (Stanford Sentiment Treebank) và Amazon Reviews cho tác vụ phân tích cảm xúc. Trong quá trình huấn luyện, lựa chọn chỉ số validation loss thấp nhất trên tập dữ liệu validation để chọn ra mô hình tốt nhất phục vụ cho quá trình thử nghiệm.

B. Experimental Details

Supervised SimCSE: Nhóm thực hiện huấn luyện mô hình có giám sát với tập dữ liệu được tạo từ tập *train*, theo định dạng đã mô tả ở phần trên (mỗi mẫu gồm 2 câu cùng cấp độ và 1 câu ở cấp độ liền kề được chọn ngẫu nhiên, đảm bảo không trùng lặp và dữ liệu được cân bằng giữa các cấp độ). Chi tiết như sau:

- **Dữ liệu:** Khoảng 300,000 mẫu với bộ dữ liệu SST và 30,000 mẫu với bộ dữ liệu amazon-reviews. *Ví dụ về bộ 3 ở bảng II dưới đây:*

Sentence	Label
It would take a complete moron to foul up a screen adaptation of Oscar Wilde's classic satire.	0
An uncomfortable movie, suffocating and sometimes almost senseless, The Grey Zone does have a center, though a morbid one.	0
This isn't a new idea.	1

Bảng II
BỘ 3 CÂU TẠO TỪ SST TRAINSET

- **Thời gian huấn luyện:** 2 giờ cho mỗi *epoch*, tổng cộng 3 *epochs*.
- **Siêu tham số:** `learning rate` = 1e-5, `temperature` = 0.05.

Unsupervised SimCSE: Nhóm thực hiện huấn luyện không giám sát với tập dữ liệu kết hợp giữa tập *train* và tập *sampled-wiki*. Sự kết hợp này nhằm đạt mục đích mô hình có thể vừa học được kiến thức về cảm xúc từ tập *train* vừa có thể học được kiến thức tổng quát từ tập *sampled-wiki*. Chi tiết như sau:

- **Thời gian huấn luyện:** 2 giờ 45 phút cho mỗi *epoch*, tổng cộng 3 *epochs*.
- **Siêu tham số:** `dropout rate` = [0.1, 0.3], `learning rate` = 1e-5, `temperature` = 0.05.

***Lưu ý:** Mô hình gốc sau khi được huấn luyện với phương pháp SimCSE sẽ được sử dụng kết hợp với MLP (đã mô tả ở phần *Model*) để tiếp tục huấn luyện cho tác vụ phân tích cảm xúc.

SMART: Nhóm áp dụng phương pháp regularization SMART (cách cài đặt được phát triển từ Github repository này [3]) trong quá trình huấn luyện, kết hợp với `GradScaler` và `autocast` (công cụ tối ưu thời gian huấn luyện thuộc thư viện

[3]https://github.com/jelc53/nlp-minibert

`torch.cuda.amp` *Automatic Mixed Precision*). Các siêu tham số cụ thể:

- **Adversarial Regularization:** `epsilon` = 1e-5, `lambda` = [1, 3, 5], `eta` = 1e-3, `sigma` = 1e-5.
- **Bregman Proximal Point:** `beta` = 0.995, `mu` = 1.

Thời gian huấn luyện gồm 10 *epochs*, với mỗi *epoch* kéo dài khoảng 20 phút.

Trong nghiên cứu này, tất cả các phương pháp huấn luyện được triển khai cho các tác vụ phân loại cảm xúc đều dựa trên việc sử dụng hàm mất mát ***cross-entropy*** *[17], một tiêu chuẩn phổ biến trong các bài toán phân loại đa lớp. Bên cạnh đó, quá trình tối ưu hóa mô hình được thực hiện bằng thuật toán* ***Adam*** *[18], một phương pháp tối ưu hóa tiên tiến, kết hợp giữa Momentum và Adaptive Learning Rate, nhằm đảm bảo tốc độ hội tụ nhanh và độ ổn định cao trong quá trình huấn luyện.*

VI. RESULTS

Nhóm thực hiện đánh giá hiệu năng của mô hình baseline so với những cách thức mở rộng (huấn luyện học tương phản có giám sát và không giám sát, phương pháp regularization SMART) trên tập dữ liệu test với độ đo accuracy và F1 (%). Kết quả thu được ở bảng III.

Model	**SST**		**amazon-reviews**	
	Acc	**F1**	**Acc**	**F1**
Baseline	52.85	50.85	57.14	57.22
SMART	**54.34**	**53.77**	**58.42**	**57.26**
Sup-CSE	53.26	52.75	56.9	57.05
Sup-CSE + SMART	52.67	52.15	56.6	54.32
Unsup-CSE	52.67	51.89	55.27	56.23
Unsup-CSE + SMART	52.94	52.10	55.73	55.21

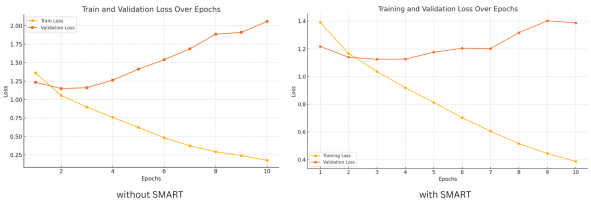
Bảng III
BẢNG KẾT QUẢ THỬ NGHIỆM

VII. ANALYSIS AND DISCUSSION

Mô hình cơ bản BERT-base, như kỳ vọng, đạt kết quả gần như thấp nhất, tái khẳng định tầm quan trọng của các kỹ thuật fine-tuning nâng cao. Tuy nhiên, sự cải thiện không quá lớn của các cách mở rộng trên tập SST cũng đặt ra câu hỏi về những thách thức đặc thù của dữ liệu này đối với các phương pháp tối ưu hóa.

SimCSE: Hai phương pháp Sup-CSE và Unsup-CSE dựa trên học tương phản cũng mang lại những cải tiến nhất định so với BERT-base. Trong đó, Sup-CSE, với lợi thế sử dụng dữ liệu có nhãn, đạt hiệu suất tốt hơn Unsup-CSE. Tuy nhiên sự chênh lệch giữa hai phương pháp này là không đáng kể, có thể do đặc thù của dữ liệu không tạo ra đủ không gian để khai thác thông tin từ nhãn.

SMART: Thực nghiệm của nhóm đã cho thấy SMART là phương pháp regularization vượt trội hơn baseline và các cách mở rộng khác. Nhóm thử nghiệm cài các siêu tham số khác nhau cho phương pháp này và nhận ra mô hình hoạt động tốt hơn trên bộ tham số hợp lý mà tác giả đã đề cập đến trong bài báo gốc: *"We only observed slight differences in model performance when* $\lambda_s \in [1, 10]$, $\mu \in [1, 10]$ *and* $\epsilon \in [10^{-5}, 10^{-4}]$. *When* $\lambda_s \geq 100$, $\mu \geq 100$ *or* $\epsilon \geq 10^{-3}$, *the regularization is unreasonably strong. When* $\lambda_s \leq 0.1$, $\mu \leq 0.1$ *or* $\epsilon \leq 10^{-6}$, *the regularization is unreasonably weak."*

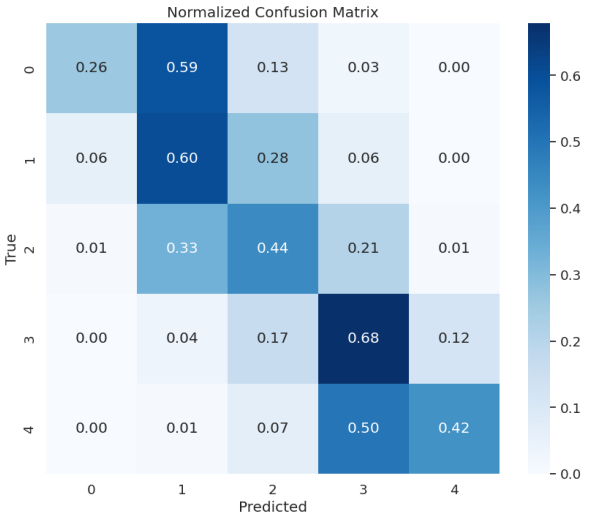


Hình 7. So sánh hiệu quả của SMART qua 10 epochs

Từ hình 7, có thể thấy phương pháp SMART giúp cân bằng giữa quá trình huấn luyện và kiểm tra. Cụ thể rằng, khi sử dụng phương pháp SMART training loss giảm chậm hơn nhưng validation loss ổn định chứ không tăng sau vài epoch đầu như khi huấn luyện không dùng phương pháp SMART (hay chính là dấu hiệu overfitting). Từ đó, phương pháp cải thiện khả năng tổng quát và hạn chế overfitting của mô hình.

Kết hợp SimCSE và SMART: Khi kết hợp 2 phương pháp này cho kết quả cải thiện không đáng kể, thậm chí là giảm hiệu năng so với áp dụng từng phương pháp. Điều này cho thấy sự phức tạp tăng lên khi áp dụng đồng thời hai phương pháp regularization và học tương phản, khiến cho mô hình khó học được các đặc điểm tinh vi của dữ liệu.

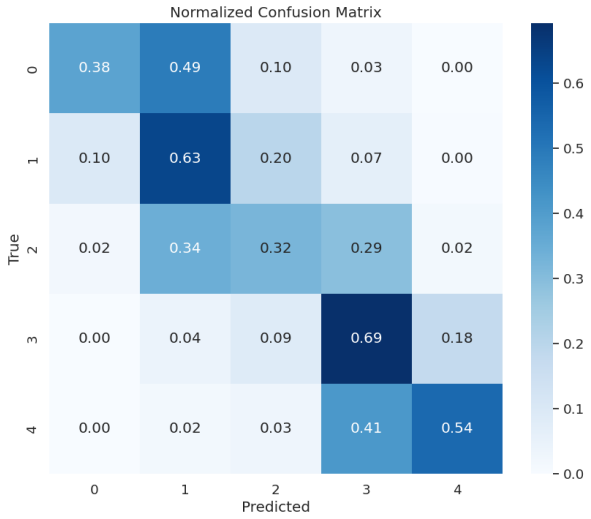
Nhận xét kỹ hơn về sự kết hợp SMART với mô hình baseline



Hình 8. Confusion Matrix cho baseline với tập dữ liệu SST

Quan sát hình 8 và 9 mô tả Confusion Matrix cho Baseline BERT và BERT áp dụng SMART trên tập

dữ liệu SST. Đầu tiên, Baseline BERT hoạt tốt ở các nhãn có số lượng dữ liệu lớn (nhãn 1 và 3), nhưng đồng thời cũng nhầm lẫn nhiều ở nhãn có số lượng dữ liệu ít hơn. Khả năng phân biệt các nhãn liền kề còn nhiều hạn chế, cũng như tại các nhãn biên như 0 và 4.



Hình 9. Confusion Matrix cho SMART với tập dữ liệu SST

Còn đối với mô hình BERT áp dụng SMART đã cải thiện accuracy của bài toán, giảm sự nhầm lẫn ở các nhãn ít xuất hiện hoặc có độ tương đồng cao nhưng chưa thực sự rõ ràng. Ngoài ra mô hình đã khắc phục khả năng phân biệt giữa cảm xúc tiêu cực và tích cực cũng như các nhãn khó ở biên. Tuy nhiên đối với nhãn 2 (trung tính) vẫn còn khá khó khăn để phân biệt với các nhãn kề (nhãn 1 và 3).

VIII. CONCLUSION

Achievements: Nhóm đã thực hiện cài thành công mô hình mở rộng của BERT-base với các phương pháp học tương phản có giám sát và không giám sát, phương pháp regularization SMART mạnh mẽ để cải thiện hiệu năng của mô hình trong tác vụ phân tích cảm xúc.

Key findings:

- Việc tinh chỉnh toàn bộ tham số của mô hình thường mang lại hiệu quả cao hơn so với việc chỉ cập nhật một phần tham số.
- SimCSE, mặc dù được áp dụng trong nhiều bối cảnh khác nhau, không cho thấy sự cải thiện đáng kể về chất lượng mô hình trong tác vụ phân tích cảm xúc.
- SMART, một phương pháp điều chuẩn (regularization), đã chứng minh khả năng cải thiện đáng kể hiệu suất mô hình. Phương pháp này hoạt động bằng cách bổ sung nhiễu vào đầu vào và giới hạn cập nhật quá mức trong quá trình huấn luyện.
- Khi kết hợp SimCSE và SMART, kết quả không nhất thiết được cải thiện. Thậm chí, độ phức tạp gia tăng trong quá trình học có thể làm giảm hiệu quả của mô hình.

Future Directions: Trong giai đoạn tiếp theo, nhóm dự định tập trung vào các hướng đi sau:

- Thử xây dựng tập train với dữ liệu các nhãn cân bằng cho quá trình huấn luyện để đảm bảo mô hình có thể học một cách tốt hơn với tất cả các nhãn(ví dụ: sinh thêm dữ liệu bằng Large Language Model ...).
- Thực hiện huấn luyện trên bộ dữ liệu lớn hơn (có thể trên bộ dữ liệu $BERT_{LARGE}$ hoặc các bộ dữ liệu tương tự), chất lượng hơn kết hợp với thời gian huấn luyện dài hơn. Điều này đặc biệt cần thiết đối với SimCSE, khi các embedding vectors yêu cầu mức độ đa dạng và đại diện cao.
- Nghiên cứu cách tối ưu hóa SimCSE để cải thiện hiệu suất trên các tác vụ yêu cầu hiểu ngữ nghĩa sâu sắc hơn, cụ thể là tác vụ phân tích cảm xúc
- Đánh giá tác động của SMART trên các kiến trúc mô hình lớn hơn như BERT-large, RoBERTa hoặc các mô hình pre-trained hiện đại khác.
- Khai thác và đánh giá hiệu suất mô hình trên các ngôn ngữ khác, đặc biệt là các ngôn ngữ có tài nguyên hạn chế.

TÀI LIỆU

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics, November 2021.

[3] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190. Association for Computational Linguistics, July 2020.

[4] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[5] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[6] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics, November 2018.

[7] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566. Association for Computational Linguistics, July 2015.

[8] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics, October 2014.

[9] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 2227–2237. Association for Computational Linguistics, June 2018.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[11] Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics, November 2019.

[12] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[14] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics, July 2018.

[15] Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14. Association for Computational Linguistics, August 2019.

[16] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995. PMLR, 2019.

[17] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.