
TRƯỜNG ĐẠI HỌC PHENIKAA
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Nhóm 8

**ỨNG DỤNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN PHÂN LOẠI ĐÁNH
GIÁ KHÁCH HÀNG TRÊN GOOGLE PLAY STORE**

Nguyễn Hoàng Dương, 21013110, 21013110@st.phenikaa-uni.edu.vn

Nguyễn Minh Đức, 21010602, 21010602@st.phenikaa-uni.edu.vn

Trần Mạnh Cường, 21011584, 21011584@st.phenikaa-uni.edu.vn

Vũ Thành Đạt, 21010589, 21010589@st.phenikaa-uni.edu.vn

GVHD: Ts. Phạm Tiến Lâm

ThS. Nguyễn Văn Sơn

11/2024

BẢNG PHÂN CHIA CÔNG VIỆC

Vai trò		Tên SV1	Tên SV2	Tên SV3	Tên SV4
		Nguyễn Hoàng Dương	Nguyễn Minh Đức	Vũ Thành Đạt	Trần Mạnh Cường
Milestone	Nội dung công việc				
I	Xây dựng ý tưởng đề tài	100%			
	Xây dựng kế hoạch đề tài				
II	Thập thập dữ liệu đánh giá khách hàng	100%			
	Phân tích dữ liệu đánh giá khách hàng				
	Tìm hiểu về mô hình Bert				
	Tiền xử lí dữ liệu thô, trực quan hóa dữ liệu				
III	Triển khai mô hình Bert, đánh giá huấn luyện mô hình trên tập dữ liệu	100%			
	Thử nghiệm mô hình				
Toàn bộ dự án	Hoàn thiện dự án	100%	100%	100%	100%

Nội Dung

BẢNG PHÂN CHIA CÔNG VIỆC.....	2
LỜI CẢM ƠN.....	5
NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN.....	6
LỜI CAM ĐOAN.....	7
MỞ ĐẦU.....	8
1.1. Tổng quan.....	9
1.2. Cơ sở khoa học.....	10
1.2.1. Một số khái niệm cơ bản.....	11
1.2.2. Xác suất (Probability).....	11
1.2.2.5. Kỳ vọng (Expectation) và Phương sai (Variance).....	12
1.3. Các thành phần chính Của NLP liên quan đến bài toán phân loại cảm xúc.....	13
1.3.1. Tiền xử lý văn bản.....	13
1.3.2. Biểu diễn văn bản (Vectorization).....	13
1.3.3. Phân loại văn bản (Text Classification).....	14
1.4 Ứng dụng phổ biến của NLP trong phân tích đánh giá người dùng.....	14
1.4.1. Phân loại cảm xúc (Sentiment Analysis).....	14
1.4.2. Phân tích chủ đề (Topic Modeling).....	14
1.4.3. Phát hiện chủ ý của khách hàng (Intent Detection).....	15
1.4.4. Tóm tắt ý kiến (Summarization).....	15
1.4.5. Phân tích từ khóa và truy vấn (Keyword Extraction & Querying).....	15
1.4.6. Phát hiện đánh giá giả mạo (Fake Review Detection).....	15
1.4.7. Phân tích ý kiến trên các mạng xã hội và tích hợp phản hồi.....	15
CHƯƠNG II: BÀI TOÁN PHÂN LOẠI ĐÁNH GIÁ KHÁCH HÀNG.....	16
2.1. Mô tả bài toán.....	16
2.2. Mục tiêu và yêu cầu bài toán.....	16
2.3. Nguồn dữ liệu.....	17
2.4. Mô hình sử dụng.....	17
CHƯƠNG III: MÔI TRƯỜNG, THƯ VIỆN SỬ DỤNG VÀ THỰC NGHIỆM.....	19
3.1. Môi trường sử dụng.....	19
3.2. Các thư viện sử dụng.....	20
3.3. Thực nghiệm.....	20
3.3.1. Tổng quan bộ dữ liệu.....	20
3.3.2. Quy trình thực nghiệm.....	22
CHƯƠNG IV: KẾT LUẬN.....	27
TÀI LIỆU THAM KHẢO.....	28

Danh mục hình ảnh

Hình 1: Kì vọng và phương sai.....	13
Hình 1: Tập dữ liệu.....	17
Hình 2: Mô hình Bert.....	18
Hình 1: Môi trường Miniconda.....	19
Hình 2: Làm sạch dữ liệu.....	22
Hình 3: Phân bố độ dài các đánh giá theo số lượng token.....	23
Hình 4: Đánh giá khách hàng từ 1-5 sao.....	23
Hình 5: Tạo và setting DataLoader.....	24
Hình 6: Epoch 1-5.....	25
Hình 7: Epoch 6-10.....	25
Hình 8: Kết quả thực nghiệm mô hình.....	26

LỜI CẢM ƠN

Lời cảm ơn đầu tiên cho chúng tôi được gửi đến các thầy cô đã giảng dạy trong trường Đại học Phenikaa những người đã truyền dạy cho chúng tôi rất nhiều những kiến thức hay và có ích để chúng tôi có thể hoàn thiện bản thân cũng như có thêm nhiều kiến thức hơn về cuộc sống. Tiếp đến cho chúng tôi được gửi lời cảm ơn đến thầy Ts. Phạm Tiến Lâm và ThS. Nguyễn Văn Sơn với sự quan tâm, dạy dỗ, chỉ bảo tận tình chu đáo của các thầy mà chúng tôi đã có thể xây dựng và hoàn thành báo cáo của mình một cách xuất sắc nhất.

Trong quá trình học tập vừa qua, với thời gian ngắn ngủi cũng như kiến thức còn nhiều thiếu sót của chúng tôi, chúng tôi mong thầy/cô có thể bỏ qua và tạo điều kiện tốt nhất cho chúng tôi để chúng tôi hoàn thành bài báo cáo này.

Chúng tôi xin chân thành cảm ơn!

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Hà Nội, ngày ... tháng ... năm 2024

GIẢNG VIÊN HƯỚNG DẪN

(Ký và ghi rõ họ tên)

LỜI CAM ĐOAN

Đề tài này là do chúng tôi tự thực hiện dựa vào một số tài liệu và không sao chép từ tài liệu hay công trình đã có trước đó. Nếu có sao chép chúng tôi xin hoàn toàn chịu trách nhiệm.

Hà Nội, ngày ... tháng ... năm 2024

NGƯỜI VIẾT

MỞ ĐẦU

Trong thời đại công nghệ 4.0, với sự bùng nổ của thông tin trên các nền tảng trực tuyến, việc khai thác dữ liệu để hiểu rõ hơn về ý kiến của người dùng đã trở thành một nhu cầu thiết yếu đối với các doanh nghiệp. Google Play Store, nơi cung cấp hàng triệu ứng dụng di động, là một nguồn dữ liệu phong phú cho việc thu thập và phân tích phản hồi của khách hàng. Những đánh giá từ người dùng trên nền tảng này không chỉ giúp nhà phát triển nắm bắt được chất lượng và hiệu suất của ứng dụng, mà còn mang lại cái nhìn sâu sắc về sự hài lòng cũng như mong đợi của khách hàng.

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực của trí tuệ nhân tạo cho phép máy tính hiểu và phân tích ngôn ngữ tự nhiên. Việc ứng dụng NLP trong phân loại đánh giá khách hàng có thể giúp tự động hóa quá trình phân tích hàng triệu đánh giá, từ đó phân loại ý kiến theo các mức độ tích cực, tiêu cực hoặc trung tính. Báo cáo này sẽ tập trung vào việc triển khai các kỹ thuật xử lý ngôn ngữ tự nhiên để phân loại đánh giá khách hàng trên Google Play Store, giúp doanh nghiệp có cái nhìn toàn diện hơn về trải nghiệm người dùng và điều chỉnh chiến lược phát triển một cách hiệu quả.

Trong phạm vi môn học này chúng em xin được trình bày về bài toán phân loại đánh giá khách hàng dựa vào NLP. Cuối cùng, mặc dù đã cố gắng rất nhiều nhưng do thời gian có hạn, khả năng dịch và hiểu và tài liệu chưa tốt nên nội dung đồ án này không thể tránh khỏi những thiếu sót, em rất mong được sự chỉ bảo, góp ý của các thầy cô và các bạn.

CHƯƠNG I: GIỚI THIỆU VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN

1.1. Tổng quan

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một lĩnh vực nghiên cứu và ứng dụng của trí tuệ nhân tạo liên quan đến tương tác giữa con người và máy tính thông qua ngôn ngữ tự nhiên. Mục tiêu chính của NLP là giúp máy tính hiểu, diễn giải và tạo ra ngôn ngữ tự nhiên một cách tự động.

NLP bao gồm nhiều thành phần và công nghệ khác nhau, bao gồm:

- **Xử lý ngôn ngữ tự nhiên:** Đây là quá trình xử lý và phân tích ngôn ngữ tự nhiên. Quá trình này bao gồm các bước như tách từ, phân tích cú pháp, phân loại ngữ nghĩa, trích xuất thông tin và truy vấn ngôn ngữ tự nhiên.
- **Hiểu ngôn ngữ tự nhiên:** Mục tiêu là giúp máy tính hiểu được ý nghĩa của ngôn ngữ tự nhiên. Điều này bao gồm việc nắm bắt ý nghĩa của câu, xác định ngữ cảnh, nhận dạng người nói hoặc tác giả, và hiểu ý đồ của người dùng.
- **Tạo ngôn ngữ tự nhiên:** Đây là quá trình tạo ra ngôn ngữ tự nhiên từ dữ liệu không phải ngôn ngữ tự nhiên. Ví dụ, quá trình này có thể bao gồm tạo câu mô tả từ dữ liệu số hoặc tạo ra văn bản từ dữ liệu có cấu trúc.
- **Dịch máy:** Dịch máy là quá trình tự động chuyển đổi văn bản từ ngôn ngữ này sang ngôn ngữ khác. Công nghệ dịch máy đã phát triển mạnh mẽ với sự xuất hiện của các mô hình học sâu như Transformers.
- **Học máy trong NLP:** Học máy đóng vai trò quan trọng trong NLP. Các phương pháp học máy, bao gồm học có giám sát và học không giám sát, được sử dụng để xây dựng các mô hình NLP có khả năng học và hiểu ngôn ngữ tự nhiên.

Ứng dụng của NLP vô cùng phong phú và đa dạng. Chúng có thể được áp dụng trong hệ thống tìm kiếm thông tin, chatbot, giao diện người-máy, phân tích ý kiến, tổ chức và tóm tắt văn bản, dò tìm tri thức và nhiều lĩnh vực khác.

NLP đã có những tiến bộ đáng kể trong thập kỷ gần đây, đặc biệt là nhờ vào sự phát triển của các mô hình học sâu và tập dữ liệu lớn. Tuy nhiên, vẫn còn nhiều thách thức trong NLP, bao gồm khả năng hiểu ngữ cảnh, xử lý ngôn ngữ không chuẩn và đa nghĩa, và đảm bảo tính công bằng và an toàn trong việc sử dụng công nghệ NLP.

1.2. Cơ sở khoa học

Cơ sở khoa học của Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) dựa trên nhiều lĩnh vực và nguyên tắc trong khoa học máy tính và ngôn ngữ học. Dưới đây là một số cơ sở khoa học quan trọng trong NLP:

1. Ngôn ngữ học : NLP dựa trên các nguyên tắc và kiến thức về ngôn ngữ học, nghiên cứu cấu trúc và chức năng của ngôn ngữ, bao gồm cú pháp, ngữ nghĩa, và ngữ âm. Ngôn ngữ học cung cấp các khái niệm và phương pháp để phân tích và hiểu ngôn ngữ tự nhiên, từ đó xây dựng nền tảng lý thuyết cho các ứng dụng NLP.

2. Xử lý ngôn ngữ tự nhiên: Xử lý ngôn ngữ tự nhiên (NLP) sử dụng các phương pháp và thuật toán để xử lý và phân tích ngôn ngữ tự nhiên. Các phương pháp này bao gồm tách từ, phân tích cú pháp, phân loại ngữ nghĩa, trích xuất thông tin, dịch máy, và nhiều công nghệ khác. NLP kết hợp ngôn ngữ học và khoa học máy tính để xây dựng các công cụ và ứng dụng liên quan đến ngôn ngữ.

3. Học máy và học sâu: Học máy chủ yếu dựa trên việc xây dựng và huấn luyện các mô hình máy tính để tự động học từ dữ liệu. Học sâu (deep learning) là một phương pháp học máy dựa trên mạng nơ-ron nhân tạo với nhiều lớp ẩn. Trong NLP, học sâu đã đạt được những tiến bộ đáng kể, đặc biệt là với sự phát triển của các mô hình Transformer như BERT, GPT và các biến thể khác, giúp cải thiện hiệu suất của các ứng dụng NLP.

4. Thống kê và xác suất: Các phương pháp và khái niệm trong thống kê và xác suất đóng vai trò quan trọng trong NLP. Các mô hình ngôn ngữ dựa trên xác suất như mô hình ngôn ngữ Markov (Markov language model) và mô hình n-gram được sử dụng để ước lượng xác suất của các câu hoặc từ. Các phương pháp thống kê cũng được sử dụng để phân tích dữ liệu ngôn ngữ và đưa ra các kết luận thống kê.

5. Xử lý dữ liệu lớn: Một trong những yếu tố quan trọng của NLP hiện đại là khả năng xử lý dữ liệu lớn. Các mô hình NLP phụ thuộc vào việc huấn

luyện trên các tập dữ liệu lớn như corpus văn bản, dữ liệu từ các mạng xã hội và các nguồn dữ liệu khác. Công nghệ xử lý dữ liệu lớn như phân tán, xử lý song song và tính toán đám mây (cloud computing) đóng vai trò quan trọng trong việc xử lý và huấn luyện các mô hình NLP.

Những cơ sở khoa học này, cùng với những tiến bộ trong công nghệ và tính toán, đã mang lại những bước tiến đáng kể trong lĩnh vực NLP, làm cho các ứng dụng NLP trở nên phổ biến và hữu ích trong thực tế.

1.2.1. Một số khái niệm cơ bản

1.2.1.1. Ngôn ngữ tự nhiên

Ngôn ngữ tự nhiên (Natural Language) là hình thức giao tiếp và truyền đạt thông tin giữa con người thông qua các phương tiện ngôn ngữ như từ ngữ, ngữ pháp và ngữ cảnh. Đây là hệ thống ngôn ngữ tự nhiên mà con người sử dụng để diễn đạt ý kiến, truyền đạt thông tin, thể hiện cảm xúc và tương tác với nhau. Đặc điểm chung của ngôn ngữ tự nhiên là khả năng phản ánh cách thức con người diễn đạt ý nghĩa và ý kiến qua ngôn từ, cú pháp, ngữ cảnh và ngữ nghĩa.

1.2.1.2. Nhập nhằng (Xử lý ngôn ngữ tự nhiên)

Nhập nhằng trong ngôn ngữ học là hiện tượng thường gặp trong giao tiếp hàng ngày, con người dễ dàng xử lý hiện tượng này nhờ vào ngữ cảnh và sự hiểu biết. Tuy nhiên, trong các ứng dụng xử lý ngôn ngữ tự nhiên, đặc biệt là dịch tự động, nhập nhằng trở thành một thách thức lớn. Ví dụ, từ "đường" trong câu "ra chợ mua cho mẹ ít đường" có thể dịch thành "road" hoặc "sugar" tùy theo ngữ cảnh. Con người dễ dàng hiểu rằng từ này nghĩa là "sugar" nhờ vào ngữ cảnh, nhưng máy tính gặp nhiều khó khăn trong việc xử lý những trường hợp nhập nhằng như vậy. Tìm ra các thuật toán hiệu quả để giải quyết nhập nhằng là một thách thức lớn trong NLP.

1.2.1.3. Dịch máy

Dịch máy là một trong những ứng dụng quan trọng của xử lý ngôn ngữ tự nhiên, dùng máy tính để dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác. Mặc dù dịch máy đã được nghiên cứu và phát triển hơn 50 năm, vẫn còn nhiều vấn đề cần nghiên cứu. Ở Việt Nam, dịch máy đã được nghiên cứu hơn 20 năm nhưng chất lượng của các sản phẩm dịch máy hiện tại vẫn còn hạn chế. Hiện nay, dịch máy được phân chia thành một số phương pháp như dịch máy dựa trên luật, dịch máy thống kê và dịch máy dựa trên ví dụ.

1.2.2. Xác suất (Probability)

1.2.2.1. Thực nghiệm và không gian mẫu

Không gian mẫu (Sự kiện cơ sở): Được ký hiệu là Ω , là tập hợp tất cả các kết quả

có thể xảy ra của một thực nghiệm.

Ví dụ:

- Tung đồng xu: $\Omega = \{\text{head}, \text{tail}\}$.
- Bầu cử: $\Omega = \{\text{yes}, \text{no}\}$.
- Tung xúc xắc: $\Omega = \{1, \dots, 6\}$.
- Xổ số: Ω có thể chứa tới hàng triệu kết quả.
- Lỗi chính tả: $\Omega = Z^*$, với Z là một bảng chữ cái và Z^* là tập hợp các chuỗi trong bảng chữ cái ($|\Omega|$ có thể rất lớn, gần bằng kích thước từ vựng).

1.2.2.2. Sự kiện (Events)

Sự kiện A là một tập hợp con của các mẫu trong không gian mẫu Ω , ký hiệu $A \subset \Omega$. Tập hợp tất cả các sự kiện là 2^Ω .

Ví dụ: Khi tung đồng xu 3 lần, không gian mẫu $\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}$.

- Sự kiện có đúng 2 lần xuất hiện "Tail": $A = \{\text{HTT}, \text{THT}, \text{TTH}\}$.
- Sự kiện tất cả là "Head": $A = \{\text{HHH}\}$.

1.2.2.3. Xác suất (Probability)

Khi thực hiện một thực nghiệm nhiều lần, số lần sự kiện A xảy ra (ký hiệu "count" c_i) được ghi lại. Tỷ lệ giữa c_i và tổng số lần thực nghiệm T_i trong dãy thực nghiệm dần tới một hằng số chưa biết, được gọi là xác suất của A (ký hiệu: $p(A)$).

1.2.2.4. Ước lượng xác suất

Để tính xác suất $p(A)$, ta có thể dùng công thức: $p(A) = c_1/T_1$, với T_1 là tổng số lần thực nghiệm trong dãy thứ i . Nếu thực hiện nhiều dãy thực nghiệm, lấy trung bình cộng của c_i/T_i .

1.2.2.5. Kỳ vọng (Expectation) và Phương sai (Variance)

- Kỳ vọng: Là tổng trọng số các giá trị của một biến ngẫu nhiên X , hay chính là giá trị trung bình của X .
- Phương sai: Là trung bình của bình phương độ lệch của biến X so với trung bình của nó, cho biết mức độ phân tán của các giá trị so với kỳ vọng.

$$E(X) = \sum_x xp(x)$$

$$Var(X) = \sum_x p(x)(x - E(x))^2$$

Hình 1: Kỳ vọng và phương sai

1.3. Các thành phần chính Của NLP liên quan đến bài toán phân loại cảm xúc

1.3.1. Tiền xử lý văn bản

- Làm sạch văn bản: Xử lý các thành phần không cần thiết như dấu câu, ký tự đặc biệt, và biểu tượng cảm xúc. Quá trình này giúp chuẩn hóa đầu vào, tránh làm nhiễu mô hình.
- Tách từ và chuẩn hóa từ vựng: Phân đoạn văn bản thành các từ (tokenization) và chuyển tất cả về dạng chữ thường. Việc chuẩn hóa này giúp giảm số lượng biến trong dữ liệu.
- Loại bỏ stop words: Loại bỏ những từ phổ biến nhưng ít mang ý nghĩa (như "và", "có", "là") để tập trung vào từ ngữ chính.
- Gốc từ (Stemming) và biến thể từ (Lemmatization): Giảm từ xuống gốc cơ bản để giảm thiểu độ phức tạp và tập trung vào ý nghĩa cốt lõi (ví dụ: "running" và "run" đều được quy về gốc từ "run").

1.3.2. Biểu diễn văn bản (Vectorization)

- Bag-of-Words (BoW): Đếm tần suất các từ xuất hiện trong văn bản và biểu diễn chúng dưới dạng vector. Tuy nhiên, BoW không xem xét đến ngữ cảnh của từ trong câu.
- TF-IDF (Term Frequency-Inverse Document Frequency): Cân nhắc độ quan trọng của các từ bằng cách kết hợp tần suất của từ trong văn bản và mức độ phổ biến của nó trên toàn bộ tập dữ liệu. TF-IDF giúp lọc ra những từ ít mang ý nghĩa.

- Embeddings (Word2Vec, GloVe, FastText): Các vector từ học được cho phép biểu diễn ngữ nghĩa của từ và ngữ cảnh của từ trong câu, giúp mô hình phân biệt ý nghĩa khác nhau của cùng một từ.
- Biểu diễn bằng BERT: Mô hình BERT tạo ra các embedding ngữ cảnh cho mỗi từ trong câu, cho phép mô hình hiểu tốt hơn ý nghĩa và sắc thái của từ trong từng ngữ cảnh khác nhau.

1.3.3. Phân loại văn bản (Text Classification)

- Mô hình học máy truyền thống: Sử dụng các thuật toán như Naive Bayes, Support Vector Machine (SVM), và Logistic Regression để phân loại văn bản dựa trên các vector từ BoW hoặc TF-IDF.
- Mô hình học sâu (Deep Learning): Các mô hình như LSTM, GRU và BERT được áp dụng cho văn bản dài và có khả năng nắm bắt ngữ cảnh và mối quan hệ giữa các từ tốt hơn. Trong phân loại cảm xúc, LSTM và BERT thường đạt kết quả tốt nhờ khả năng nắm bắt ngữ cảnh và cảm xúc trong văn bản.
- Fine-tuning trên BERT: Điều chỉnh mô hình BERT đã được huấn luyện trên kho dữ liệu lớn để thích nghi với nhiệm vụ cụ thể như phân loại cảm xúc, nhờ đó đạt hiệu suất cao mà không cần lượng dữ liệu huấn luyện quá lớn.

1.4 Ứng dụng phổ biến của NLP trong phân tích đánh giá người dùng

1.4.1. Phân loại cảm xúc (Sentiment Analysis)

Đây là ứng dụng phổ biến nhất trong phân tích đánh giá người dùng, giúp xác định cảm xúc của người đánh giá là tích cực, tiêu cực hay trung tính. Các hệ thống phân loại cảm xúc hỗ trợ doanh nghiệp nắm bắt sự hài lòng của khách hàng, phát hiện các vấn đề tiềm ẩn và cải thiện sản phẩm kịp thời.

1.4.2. Phân tích chủ đề (Topic Modeling)

Sử dụng các phương pháp như LDA (Latent Dirichlet Allocation) để xác định chủ đề chính trong đánh giá của khách hàng. Phân tích chủ đề giúp nhà phát triển

xác định những điểm quan trọng mà khách hàng đang đề cập, ví dụ như "hiệu suất", "giao diện", "giá cả".

1.4.3.Phát hiện chủ ý của khách hàng (Intent Detection)

Phát hiện ý đồ của người dùng trong đánh giá, chẳng hạn như đánh giá có mang tính đề xuất tính năng, phản ánh lỗi hoặc chỉ trích. Điều này hỗ trợ phân loại đánh giá theo mức độ ưu tiên để có hướng xử lý phù hợp.

1.4.4.Tóm tắt ý kiến (Summarization)

Tóm tắt một lượng lớn đánh giá thành các điểm chính, giúp doanh nghiệp nắm bắt ý kiến chung của khách hàng mà không cần đọc qua từng đánh giá. NLP sử dụng các kỹ thuật tóm tắt văn bản để rút gọn đánh giá thành các ý chính.

1.4.5.Phân tích từ khóa và truy vấn (Keyword Extraction & Querying)

Trích xuất từ khóa và cụm từ quan trọng từ các đánh giá để hỗ trợ việc tìm kiếm và phân tích sâu hơn về các yếu tố quan trọng với khách hàng, như "dịch vụ khách hàng", "tốc độ ứng dụng", "độ ổn định".

1.4.6.Phát hiện đánh giá giả mạo (Fake Review Detection)

NLP giúp phát hiện các đánh giá giả mạo hoặc spam bằng cách phân tích cấu trúc câu, ngữ nghĩa và hành vi người dùng, từ đó đảm bảo tính chính xác của dữ liệu đánh giá và tránh bị thao túng bởi các đánh giá không trung thực.

1.4.7.Phân tích ý kiến trên các mạng xã hội và tích hợp phản hồi

Kết hợp đánh giá từ Google Play Store với dữ liệu từ mạng xã hội hoặc các nguồn khác để có cái nhìn toàn diện hơn về phản hồi của người dùng, từ đó điều chỉnh chiến lược phát triển và cải thiện dịch vụ.

CHƯƠNG II: BÀI TOÁN PHÂN LOẠI ĐÁNH GIÁ KHÁCH HÀNG

2.1. Mô tả bài toán

Trong bối cảnh hiện đại, các ứng dụng di động ngày càng phổ biến, và cửa hàng ứng dụng như Google Play Store là nơi người dùng để lại các đánh giá trực tiếp sau khi sử dụng. Những đánh giá này không chỉ giúp người dùng tiềm năng quyết định tải về ứng dụng mà còn cung cấp thông tin quý giá cho nhà phát triển về hiệu suất và trải nghiệm ứng dụng. Tuy nhiên, không phải tất cả các đánh giá đều tích cực và nhiều trong số đó mang tính chất tiêu cực hoặc trung tính, phản ánh sự không hài lòng hoặc đề xuất cải tiến.

Bài toán phân loại đánh giá khách hàng hướng đến việc tự động phân tích và sắp xếp các đánh giá này dựa trên cảm xúc tích cực, tiêu cực, trung tính hoặc mức độ hài lòng, tạo điều kiện cho việc tổng hợp phản hồi từ người dùng dễ dàng và hiệu quả hơn. Ngoài ra, việc phát hiện đánh giá giả mạo hoặc không trung thực cũng đóng vai trò quan trọng trong việc bảo vệ tính chân thực của dữ liệu phản hồi.

2.2. Mục tiêu và yêu cầu bài toán

Mục tiêu chính của bài toán phân loại đánh giá khách hàng bao gồm:

- Phân loại cảm xúc đánh giá: Xác định loại cảm xúc của mỗi đánh giá (tích cực, tiêu cực, trung tính) để từ đó đưa ra bức tranh toàn cảnh về mức độ hài lòng của người dùng.
- Phân loại theo mức độ hài lòng: Sắp xếp các đánh giá theo mức độ hài lòng, chẳng hạn như từ 1-5 sao, nhằm cung cấp các báo cáo chi tiết về mức độ hài lòng của người dùng, giúp đội ngũ phát triển cải thiện sản phẩm dựa trên phản hồi cụ thể.
- Phát hiện đánh giá giả mạo: Đánh giá giả mạo thường không mang tính trung thực và có thể ảnh hưởng đến quyết định của người dùng khác. Việc phát hiện và loại bỏ các đánh giá này giúp duy trì tính minh bạch và uy tín của ứng dụng.

2.3. Nguồn dữ liệu

Nhóm em thu thập dữ liệu từ nền tảng Kaggle, dữ liệu bao gồm hơn 12000 đánh giá từ các ứng dụng khác nhau trên Google Play Store từ người dùng chứa đầy đủ các thông tin cần thiết như:

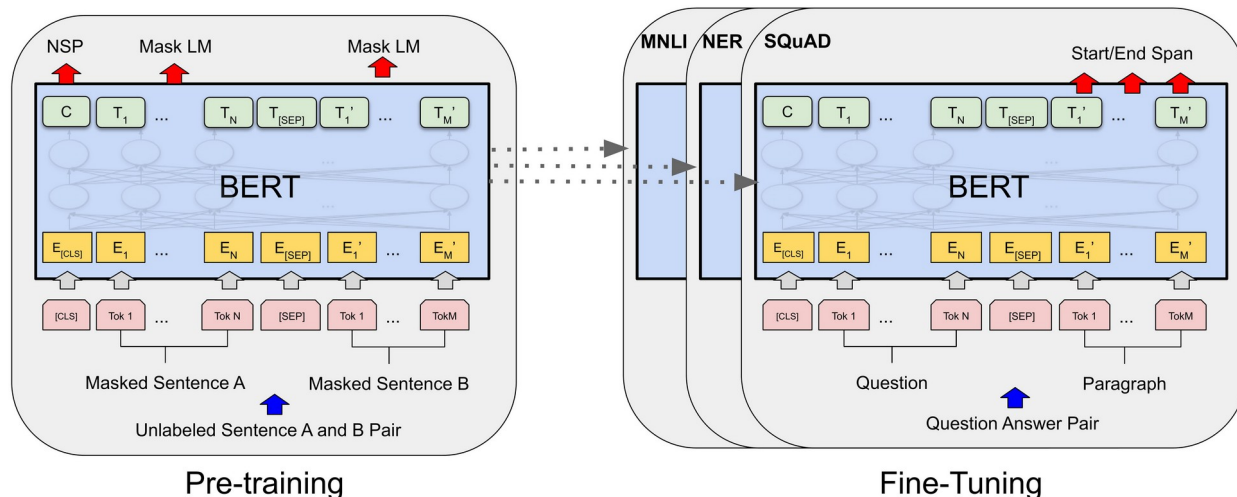
- **Nội dung đánh giá:** Các nhận xét của người dùng về ứng dụng, phản ánh cảm xúc và ý kiến của họ.
- **Xếp hạng (số sao):** Thể hiện mức độ hài lòng của người dùng, thường từ 1 đến 5 sao.
- **Thông tin về người đánh giá:** Một số thông tin cơ bản về người dùng để xác định tính nhất quán và phát hiện đánh giá giả mạo.

```
1 reviewId,userName,userImage,content,score,thumbsUpCount,reviewCreatedVersion,at,replyContent,repliedAt,sortOrder,appId
2 gp:A0qpTOEHZuqSqqWnakRgv-9ABYdaJFUB0WugPGh-SG-fgH355YH_t7J2q4xYo6ZzN3Mc7iSrrTV6ke8hG_fl4Q,Eric Tie,https://play-lh.googleusercontent.com/a-/A0h14GiGET2XH
3 gp:A0qpTOH0WP4IQBZ2LrdNmFy_YmpPCVrV3diEU9KGm3fAX6VG0NAZCudCQpQRRi3GLL_tr8DQzUTP1hr0YG74A,john alpha,https://play-lh.googleusercontent.com/a-/A0h14GjpfjgJ
4 gp:A0qpTOEMckJB8Iq1p-r9dPwnSYadA5BkPWTf32Z1azuUtvqA9KwdTQqNNXWZsJEhmSuYUY_LmL-OduI14j70wg,Sudhakar .S,https://play-lh.googleusercontent.com/a-/A0h14GidHU
5 gp:A0qpTOGFrUWuKgcpcje8ksZj3uwHN6tU_fd4gLVFy9z7hfgM7Gan22TJrN89NmGVEDj5o4U6W4I6sLbTx80sQw,SKGflorida@bellsouth.net DAVID S,https://play-lh.googleusercontent
6 gp:A0qpTOHLS7DW8wmdFzTkhWxuqFkdNQtkHm06Pt9jhZEQ0Q2rDzcc9WMABIXNu0pIJOhiFrA4uhMOLq1ZIMKQA,Louann Stoker,https://play-lh.googleusercontent.com/-pBcY_Z-qfB
7 gp:A0qpTOEVLpSba6k8rLDmk-WrEoJea98KURIGYwodJ-FA0Nx09WflbZ3CLazVJJDAAp3dVvd1g,Jon Clemens,https://play-lh.googleusercontent.com/-q6L1fx0d77w/
8 gp:A0qpTOGhb-APKXNNFYLI0uWgq1AGW6bQp5aYYxSHvJ_Col 1:reviewid_pWFPNb10urenFAS-p4fxP4490-C0B8A,Gale W,https://play-lh.googleusercontent.com/a-/A0h14GgZuh05SN1
9 gp:A0qpTOECxvv_c0CK3G5tHjXs6SjbzD650Q3og02p-q1rR1qy8mbJr6yE585uRIZ9xQPOMiP6V6LodVmiJm3aw,No One,https://play-lh.googleusercontent.com/a-/A0h14GhY0I-W2AG
10 gp:A0qpTOHVgr7_q0hORsfPwnVCqX-n98dJ1Ksrps6q52pgv44BNU_oSk1oprSIE9CwfYefD-kFIRjzGo4F4CRMYQ,I Dewa Gede Nopi Ariana,https://play-lh.googleusercontent.com/a
```

Hình 1: Tập dữ liệu

2.4. Mô hình sử dụng

Phương pháp sử dụng trong bài toán phân loại đánh giá khách hàng trên Google Play Store là mô hình BERT (Bidirectional Encoder Representations from Transformers). BERT là một mô hình học sâu tiên tiến được phát triển bởi Google, đặc biệt mạnh mẽ trong việc xử lý các bài toán ngôn ngữ tự nhiên, bao gồm phân loại cảm xúc, trả lời câu hỏi và dịch máy.



Hình 2: Mô hình Bert

Lí do chọn sử dụng:

Về mặt lý thuyết, các kỹ thuật khác như Word2vec, FastText hay Glove cũng tìm ra đại diện của từ thông qua ngữ cảnh chung của chúng. Tuy nhiên, những ngữ cảnh này là đa dạng trong dữ liệu tự nhiên. Ví dụ các từ như "con chuột" có ngữ nghĩa khác nhau ở các ngữ cảnh khác nhau như "Con chuột máy tính này thật đẹp!!" và "con chuột này to thật." Trong khi các mô hình như Word2vec, fastText tìm ra 1 vector đại diện cho mỗi từ dựa trên 1 tập ngữ liệu lớn nên không thể hiện được sự đa dạng của ngữ cảnh. Việc tạo ra một biểu diễn của mỗi từ dựa trên các từ khác trong câu sẽ mang lại kết quả ý nghĩa hơn nhiều. Như trong trường hợp trên ý nghĩa của từ con chuột sẽ được biểu diễn cụ thể dựa vào phần trước hoặc sau nó trong câu. Nếu đại diện của từ "con chuột" được xây dựng dựa trên những ngữ cảnh cụ thể này thì ta sẽ có được biểu diễn tốt hơn.

BERT mở rộng khả năng của các phương pháp trước đây bằng cách tạo các biểu diễn theo ngữ cảnh dựa trên các từ trước và sau đó để dẫn đến một mô hình ngôn ngữ với ngữ nghĩa phong phú hơn.

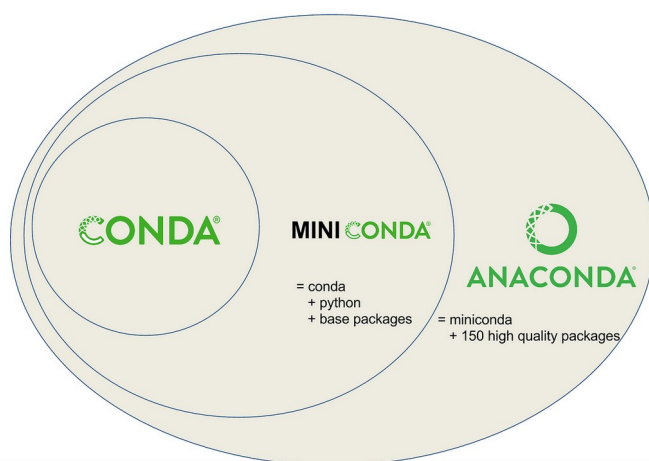
Ngoài ra BERT có khả năng fine-tune (tinh chỉnh) dễ dàng cho nhiều tác vụ NLP khác nhau. Với mô hình BERT đã được huấn luyện trước trên một lượng dữ liệu lớn, ta chỉ cần tinh chỉnh mô hình này trên bộ dữ liệu của bài toán cụ thể (như phân loại cảm xúc trong đánh giá) mà không cần phải huấn luyện lại từ đầu. Điều này giúp tiết kiệm thời gian và tài nguyên tính toán, đồng thời đạt được kết quả nhanh chóng.

CHƯƠNG III: MÔI TRƯỜNG, THƯ VIỆN SỬ DỤNG VÀ THỰC NGHIỆM

3.1. Môi trường sử dụng

Môi trường phát triển cho bài toán phân loại cảm xúc đánh giá khách hàng trên Google Play Store sử dụng Miniconda cùng với ide là VSCode.

Miniconda là một phiên bản nhẹ của Anaconda, giúp quản lý môi trường ảo và cài đặt các thư viện cần thiết cho các dự án Python một cách hiệu quả. Miniconda chỉ cài đặt conda và không bao gồm các thư viện mặc định như Anaconda, giúp giảm bớt dung lượng và tùy chỉnh môi trường dễ dàng.



Hình 1: Môi trường Miniconda

Một số điểm nổi bật của Miniconda:

- **Quản lý môi trường ảo:** Conda cho phép tạo và quản lý các môi trường ảo, mỗi môi trường có thể cài đặt các thư viện và phiên bản Python riêng biệt mà không ảnh hưởng đến các môi trường khác. Điều này giúp quản lý các phụ thuộc của dự án tốt hơn và tránh xung đột giữa các thư viện.
- **Cài đặt thư viện nhanh chóng:** Cài đặt các thư viện dễ dàng, đặc biệt là các thư viện có yêu cầu về phụ thuộc (dependencies) như TensorFlow, PyTorch hay BERT, mà không gặp phải các vấn đề thường gặp khi sử dụng pip.
- **Tiết kiệm dung lượng:** So với Anaconda, Miniconda chiếm ít dung lượng hơn và giúp bạn chỉ cài đặt các công cụ cần thiết cho dự án của mình.

3.2. Các thư viện sử dụng

NumPy: là thư viện cơ bản và phổ biến để xử lý các phép toán số học và mảng (arrays) trong Python. Nó hỗ trợ các cấu trúc dữ liệu như mảng đa chiều (ndarray), các phép toán vector hóa, và các phép toán tuyến tính, đại số ma trận.

Pandas: xử lý và phân tích dữ liệu dạng bảng (dataframe). Thư viện này giúp thao tác với dữ liệu, bao gồm các tác vụ như lọc, nhóm, sắp xếp và xử lý các giá trị thiếu (missing values).

Seaborn: dùng trực quan hóa dữ liệu, được xây dựng trên Matplotlib. Nó giúp tạo ra các biểu đồ thống kê đẹp mắt và dễ đọc. Seaborn hỗ trợ nhiều loại biểu đồ như scatter plot, bar plot, heatmap, v.v.

Sklearn (scikit-learn): scikit-learn, thường được gọi là sklearn, là một thư viện phổ biến trong Python dùng cho machine learning và data mining. Nó cung cấp các công cụ cho việc xây dựng và huấn luyện các mô hình học máy, thực hiện các quy trình tiền xử lý dữ liệu và đánh giá các mô hình.

PyTorch: là thư viện phổ biến cho việc phát triển và huấn luyện các mô hình học sâu (deep learning).

Matplotlib.pyplot: dùng để tạo và hiển thị đồ thị và biểu đồ. Nó giúp bạn trực quan hóa dữ liệu và tùy chỉnh hình dạng, kiểu đồ thị theo ý muốn.

Transformers: cung cấp các mô hình tiền huấn luyện (pre-trained models) như BERT, GPT, T5, v.v., giúp tiết kiệm thời gian huấn luyện và đạt được hiệu quả cao trong các tác vụ NLP. BertModel là mô hình BERT cơ bản, trong khi BertTokenizer dùng để mã hóa (tokenize) văn bản đầu vào thành các chỉ mục mà mô hình có thể xử lý.

3.3. Thực nghiệm

3.3.1. Tổng quan bộ dữ liệu

Tập dữ liệu đánh giá khách hàng bao gồm 12,495 dòng, mỗi dòng tương ứng với một đánh giá của người dùng về ứng dụng. Dưới đây là các đặc điểm chi tiết của tập dữ liệu này:

- ◆ **Cấu trúc của mỗi dòng dữ liệu:**

- Thông tin về người dùng:

- `userName`: Tên người dùng đã thực hiện đánh giá.
- `userImage`: Liên kết đến ảnh đại diện của người dùng.
- Nội dung đánh giá:
 - `content`: Phần văn bản chính chứa ý kiến của người dùng, phản ánh trải nghiệm cá nhân.
 - `score`: Điểm đánh giá từ 1 đến 5, có thể sử dụng làm nhãn cảm xúc cho các bài toán phân tích cảm xúc (chẳng hạn, 1-2 điểm có thể coi là tiêu cực, 4-5 là tích cực, và 3 là trung tính).
 - `thumbsUpCount`: Số lượt thích đánh giá từ những người dùng khác.
- Thông tin khác:
 - `reviewCreatedVersion`: Phiên bản ứng dụng khi đánh giá được viết.
 - `at`: Thời điểm đánh giá được tạo ra.
 - `replyContent` và `repliedAt`: Phản hồi của nhà phát triển (nếu có), cho thấy các nhà phát triển đã trả lời như thế nào với phản hồi của người dùng.

◆ Phân bố của các nhãn điểm đánh giá (score):

- Số lượng đánh giá ở mỗi mức điểm:
 - 5 sao: 2,879 đánh giá
 - 4 sao: 2,775 đánh giá
 - 3 sao: 1,991 đánh giá
 - 2 sao: 2,344 đánh giá
 - 1 sao: 2,506 đánh giá
- Phân bố này khá cân bằng giữa các mức độ đánh giá, tuy nhiên, có thể thấy số lượng đánh giá tích cực (4-5 sao) và tiêu cực (1-2 sao) là tương đối cao, trong khi đánh giá trung tính (3 sao) ít hơn.

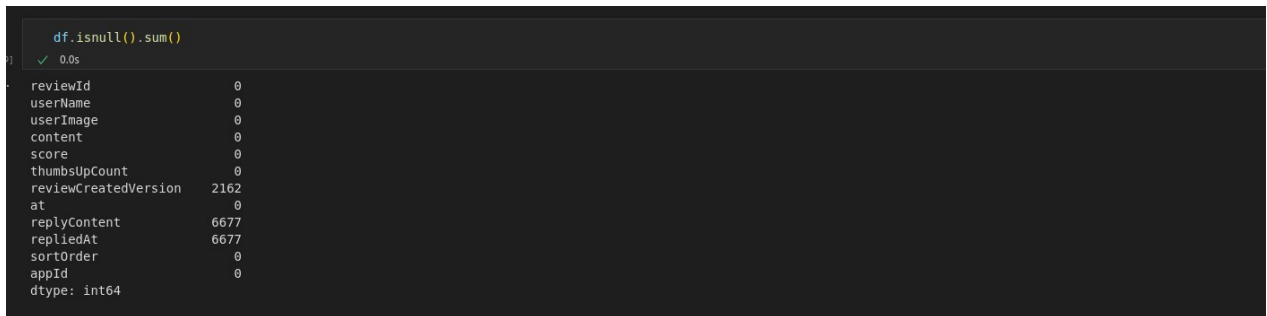
◆ Khả năng ứng dụng trong các bài toán NLP:

- Phân tích cảm xúc: Tập dữ liệu này có thể được sử dụng để phân tích cảm xúc tự động, với các điểm số đánh giá làm nhãn cảm xúc. Dữ liệu có thể được tiền xử lý như chuyển thành chữ thường, loại bỏ các ký tự đặc biệt và dừng từ (stop words), và tokenization (tách từ).
- Phân cụm: Nếu không có nhãn cảm xúc (ví dụ, chưa gán nhãn dựa trên điểm số), dữ liệu này có thể áp dụng cho bài toán phân cụm (clustering) để tìm ra các nhóm đánh giá có nội dung tương đồng. Các phương pháp như TF-IDF hoặc Word Embeddings (nhúng từ) sẽ hữu ích trong việc chuyển văn bản thành vector số học, giúp phân cụm hiệu quả.

3.3.2. Quy trình thực nghiệm

Tiền xử lí dữ liệu:

Xóa các dữ liệu bị thiếu và tách dữ liệu theo cột cụ thể (`reviewText` và `sentiment`) để phù hợp với đầu vào mô hình. Thông thường, bước này giúp chuẩn bị dữ liệu đầu vào cho mô hình BERT.



```
df.isnull().sum()
0.0s
reviewId      0
userName      0
userImage     0
content       0
score         0
thumbsUpCount 0
reviewCreatedVersion 2162
at            0
replyContent  6677
repliedAt     6677
sortOrder     0
appId         0
dtype: int64
```

Hình 2: Làm sạch dữ liệu

Trực quan hóa label thể hiện điểm đánh giá khách hàng

Tokenization:

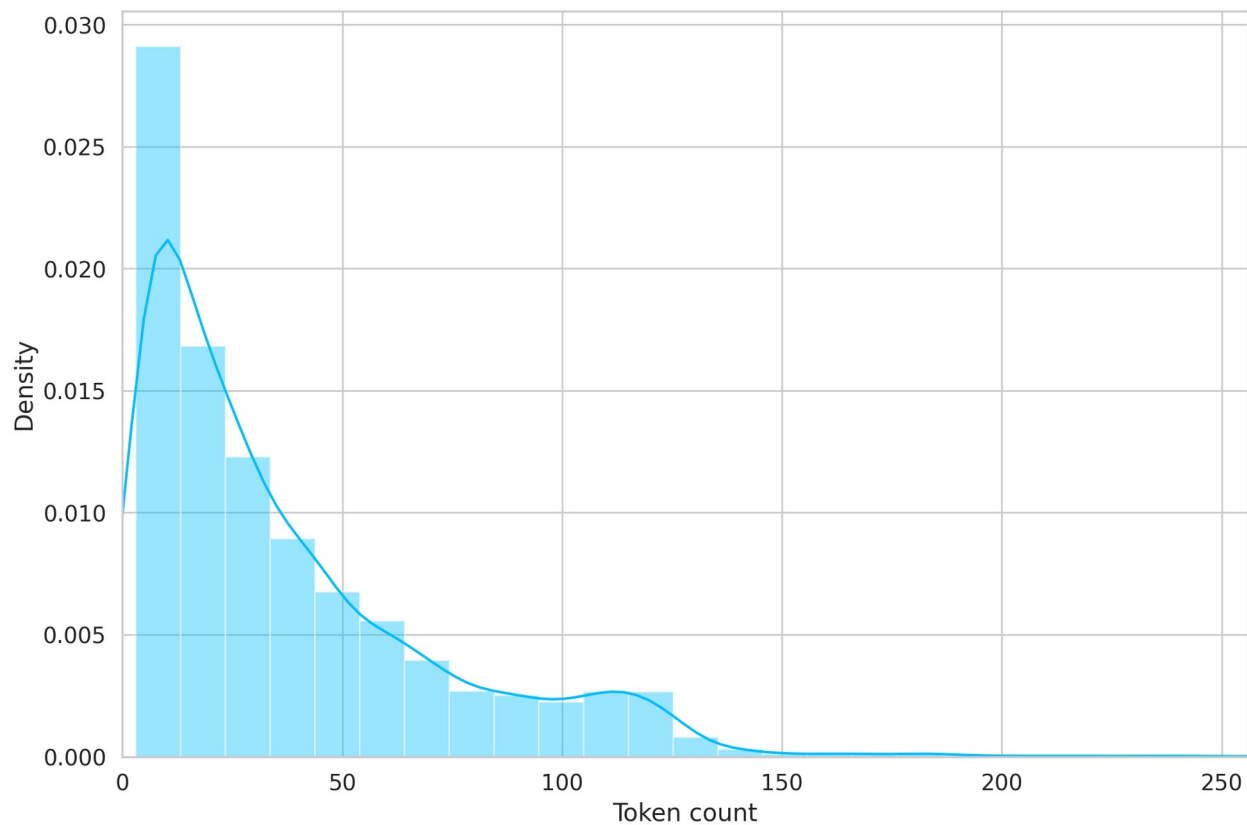
Tokenization là bước biến đổi văn bản thành các "token" – thường là các từ hoặc cụm từ đơn lẻ để xử lý bằng mô hình BERT.

- Sử dụng tokenizer của BERT: Tokenizer của BERT chuyển văn bản thành các ID số tương ứng với từ vựng mà BERT đã học được. Mỗi câu được chuyển thành một chuỗi số, biểu diễn cho các từ, bắt đầu bằng token `[CLS]` và kết thúc bằng `[SEP]`.

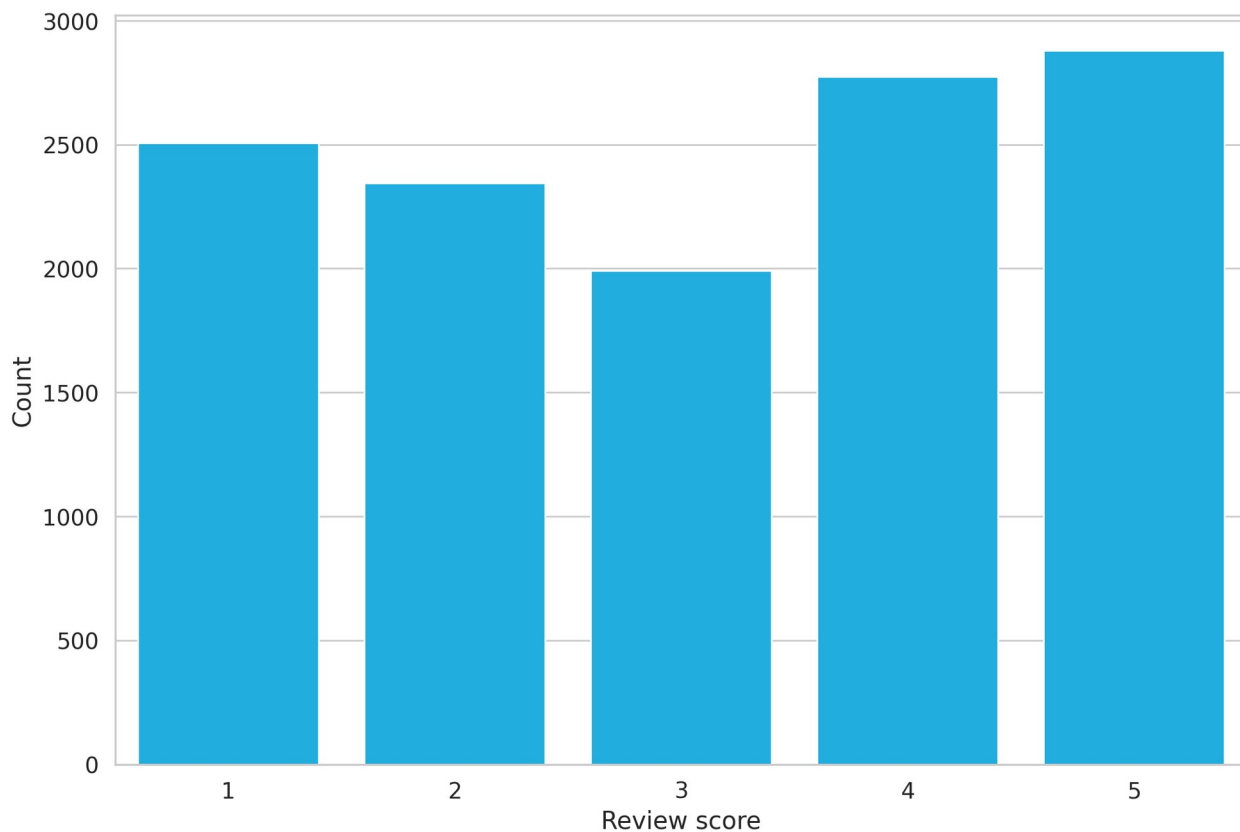
- Padding và Truncation:

Các câu có độ dài khác nhau được padding để có cùng chiều dài, giúp mô hình xử lý dễ dàng hơn.

Các câu quá dài sẽ được truncation để giảm về độ dài tối đa cho phép.



Hình 3: Phân bố độ dài các đánh giá theo số lượng token



Hình 4: Đánh giá khách hàng từ 1-5 sao

Chia dữ liệu thành các tập train, test

Sử dụng `train_test_split` để chia dữ liệu, khoảng 80% cho huấn luyện và 20% cho kiểm tra.

Tạo DataLoader

Sử dụng `torch.utils.data.DataLoader` để chuẩn bị dữ liệu thành các batch, giúp tăng tốc độ huấn luyện và đảm bảo mô hình hoạt động hiệu quả.

```
def create_data_loader(df, tokenizer, max_len, batch_size):
    ds = GPreViewDataset(
        reviews=df.content.to_numpy(),
        targets=df.sentiment.to_numpy(),
        tokenizer=tokenizer,
        max_len=max_len
    )

    return DataLoader(
        ds,
        batch_size=batch_size,
        num_workers=0
    )

BATCH_SIZE = 16
train_data_loader = create_data_loader(df_train, tokenizer, MAX_LEN, BATCH_SIZE)
val_data_loader = create_data_loader(df_val, tokenizer, MAX_LEN, BATCH_SIZE)
test_data_loader = create_data_loader(df_test, tokenizer, MAX_LEN, BATCH_SIZE)
```

Hình 5: Tạo và setting DataLoader

Xây dựng model

Sử dụng BERT cho bài toán phân loại cảm xúc, với đầu ra phân loại (binary classification), dựa trên các đánh giá tích cực, tiêu cực hoặc trung tính.

Phân tích kết quả huấn luyện và đánh giá


```

Epoch 1/10
-----
Train loss 0.6976516467809677 accuracy 0.7262905162064827
Val loss 0.5941510389122782 accuracy 0.7678142514011208

Epoch 2/10
-----
Train loss 0.4929004891037941 accuracy 0.8075230092036815
Val loss 0.6335398396359214 accuracy 0.7742193755004003

Epoch 3/10
-----
Train loss 0.3562534824281931 accuracy 0.872248899559824
Val loss 0.7233484569418279 accuracy 0.7566052842273818

Epoch 4/10
-----
Train loss 0.2579423754528165 accuracy 0.9181672669067628
Val loss 0.9641544645109886 accuracy 0.7421937550040032

Epoch 5/10
-----
Train loss 0.1902366361485794 accuracy 0.9447779111644659
Val loss 1.1575154943651036 accuracy 0.7582065652522018

```

Hình 6: Epoch 1-5

```

Epoch 6/10
-----
Train loss 0.14030329982889816 accuracy 0.9600840336134454
Val loss 1.3756306885352618 accuracy 0.7534027221777422

Epoch 7/10
-----
Train loss 0.1070378261920996 accuracy 0.9714885954381753
Val loss 1.502558264598439 accuracy 0.7429943955164131

Epoch 8/10
-----
Train loss 0.09141527631748468 accuracy 0.9753901560624251
Val loss 1.5690113146895472 accuracy 0.7502001601281024

Epoch 9/10
-----
Train loss 0.07448084465935827 accuracy 0.9806922769107643
Val loss 1.63927910503448 accuracy 0.7421937550040032

Epoch 10/10
-----
Train loss 0.0655395959211979 accuracy 0.9821928771508605
Val loss 1.693993733426603 accuracy 0.7461969575660528

```

Hình 7: Epoch 6-10

Train loss: Giảm dần theo từng epoch, từ 0.697 ở epoch đầu tiên xuống còn 0.065 ở epoch 10. Điều này cho thấy mô hình đang học dần dần và giảm thiểu lỗi trong suốt quá trình huấn luyện.

Train accuracy: Tăng dần từ 72.6% ở epoch đầu tiên lên tới 98.2% ở epoch 10. Đây là dấu hiệu mô hình đang học tốt hơn trên tập huấn luyện.

Val loss: Bắt đầu tăng từ epoch 3 trở đi, đạt đến 1.69 ở epoch cuối cùng.

Val accuracy: Ban đầu tăng nhẹ, đạt đỉnh tại epoch 2 với 77.4%, nhưng sau đó giảm dần và dao động khoảng 74-75%.

Đánh giá overfitting: Sau một số epoch đầu tiên, val loss tăng dần, còn val accuracy không cải thiện. Điều này cho thấy mô hình vẫn còn bị overfitting và cần được cải thiện.

Đề xuất thêm kỹ thuật regularization (như dropout) để hạn chế overfitting, giúp mô hình học các đặc điểm tổng quát hơn thay vì ghi nhớ dữ liệu huấn luyện.

Kết quả thu được

Thu được 1 mô hình đánh giá tổng quan khách hàng dựa trên review text với accuracy lên đến hơn 98%.

Với nội dung "need to speed up", người dùng dường như đang đưa ra một nhận xét về tốc độ của dịch vụ/sản phẩm, nhưng không quá tích cực hay tiêu cực. Do đó, dự đoán "neutral" là hợp lý. Kết quả này cho thấy mô hình có khả năng nhận diện các cảm xúc khá tốt và chính xác, có thể cải thiện với tập dataset lớn hơn trong tương lai.

```
encoded_review = tokenizer.encode_plus(  
    review_text,  
    max_length=MAX_LEN,  
    add_special_tokens=True,  
    return_token_type_ids=False,  
    pad_to_max_length=True,  
    return_attention_mask=True,  
    return_tensors='pt',  
)  
✓ 0.0s  
  
input_ids = encoded_review['input_ids'].to(device)  
attention_mask = encoded_review['attention_mask'].to(device)  
  
output = model(input_ids, attention_mask)  
_, prediction = torch.max(output, dim=1)  
  
print(f'Review text: {review_text}')print(f'Sentiment : {class_names[prediction]}')✓ 0.0s  
  
Review text: need to speed up  
Sentiment : neutral
```

Hình 8: Kết quả thực nghiệm mô hình

CHƯƠNG IV: KẾT LUẬN

Trong báo cáo này, nhóm đã thực hiện phân tích và phân loại cảm xúc của các đánh giá khách hàng sử dụng mô hình BERT. Việc áp dụng BERT – một mô hình ngôn ngữ mạnh mẽ dựa trên kiến trúc transformer, đã mang lại hiệu quả cao trong việc xử lý và hiểu ngữ nghĩa của ngôn ngữ tự nhiên. Mục tiêu chính của dự án là xác định cảm xúc của khách hàng thông qua các đánh giá của họ, từ đó hỗ trợ doanh nghiệp nắm bắt được mức độ hài lòng và nhận diện các vấn đề cần cải thiện trong dịch vụ hoặc sản phẩm của mình.

Trong quá trình thực nghiệm, chúng tôi đã trải qua các bước từ tiền xử lý dữ liệu, mã hóa văn bản, đến huấn luyện mô hình và đánh giá kết quả. Dữ liệu đầu vào đã được xử lý một cách cẩn thận, bao gồm loại bỏ các dòng dữ liệu bị thiếu, chuẩn hóa văn bản về chữ thường, và loại bỏ các từ không mang ý nghĩa quan trọng (stop words). Chúng tôi cũng thực hiện tokenization để chuyển đổi văn bản thành các token số, phục vụ cho quá trình mã hóa của BERT.

Mô hình BERT được huấn luyện qua nhiều epoch, cho thấy sự tiến bộ rõ rệt trong các chỉ số hiệu suất qua từng epoch ban đầu. Tuy nhiên, sau khoảng 3 đến 4 epoch, mô hình bắt đầu có dấu hiệu của hiện tượng overfitting, khi độ chính xác trên tập huấn luyện tiếp tục tăng cao trong khi độ chính xác trên tập kiểm tra có xu hướng giảm. Điều này cho thấy mô hình đã bắt đầu "ghi nhớ" quá mức các đặc điểm của dữ liệu huấn luyện mà không tổng quát hóa tốt cho dữ liệu mới.

Ngoài ra, qua việc kiểm tra các đánh giá mà mô hình phân loại sai, chúng tôi nhận thấy rằng một số đánh giá chứa các từ ngữ đa nghĩa hoặc mỉa mai (sarcasm), gây khó khăn cho mô hình. Đây là một thách thức trong phân tích cảm xúc vì các mô hình ngôn ngữ hiện tại vẫn gặp giới hạn trong việc hiểu các ngữ nghĩa phức tạp và hàm ý sâu xa. Điều này mở ra hướng phát triển trong tương lai, đó là kết hợp các mô hình phân tích ngữ nghĩa chuyên sâu hoặc các kỹ thuật khác như multi-task learning để nâng cao khả năng nhận diện các cảm xúc phức tạp.

Tóm lại, nghiên cứu này đã cho thấy tiềm năng của mô hình BERT trong phân loại cảm xúc của các đánh giá khách hàng, với độ chính xác cao và khả năng xử lý ngữ nghĩa ngữ cảnh tốt. Tuy nhiên, để áp dụng vào thực tế, cần có những cải tiến và điều chỉnh để tăng khả năng tổng quát hóa của mô hình, cũng như xử lý các đánh giá phức tạp hơn. Trong tương lai, việc mở rộng dữ liệu huấn luyện, kết hợp với các kỹ thuật tinh chỉnh chuyên sâu và các mô hình tiên tiến hơn sẽ là bước tiếp theo để nâng cao hiệu quả của hệ thống phân loại cảm xúc từ văn bản tự nhiên, từ đó mang lại giá trị thực tiễn cao hơn cho các doanh nghiệp.

TÀI LIỆU THAM KHẢO

- [1]<https://www.kaggle.com/code/prakharrathi25/sentiment-analysis-using-bert/notebook>
- [2]<https://viblo.asia/p/bert-buoc-dot-pha-moi-trong-cong-nghe-xu-ly-ngon-ngu-tu-nhien-cua-google-RnB5pGV7IPG>
- [3]https://medium.com/@manjindersingh_10145/sentiment-analysis-with-bert-using-huggingface-88e99deec9a