**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

-------------------------------------------



**Phương Pháp Nghiên Cứu Khoa Học**

**PHÂN TÍCH CẢM XÚC CỦA KHÁCH HÀNG TRONG VIỆC ĐÁNH GIÁ SẢN PHẨM SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY**

**Giảng Viên Hướng Dẫn : TS. Đỗ Như Tài**

**Sinh Viên Thực Hiện**

**Diệp Thụy An 3122410001**

**Đỗ Mai Anh 3122400006**

**Cao Tiến Cường 3122410043**

**Võ Hoàng Phúc Hy 3123410142**

**Tháng 4/2025 - Thành Phố Hồ Chí Minh**

**MỤC LỤC**

[**Chương I: Tổng Quan Vấn Đề 4**](#_9r6cm0fhb10n)

[**Chương II: Lược khảo tài liệu 6**](#_cmeq6k27cs6m)

[**III: Phương Pháp Nghiên Cứu 9**](#_sigisqtqbeyo)

[**3.3. Review length analysis (độ dài các đánh giá) 12**](#_n2kdoh7o2ec3)

[**3.4. Correlation analysis (phân tích tương quan) 14**](#_61nahhxaro5e)

[**3.5. Conclusion (kết luận) 15**](#_j9475q6q7noo)

[**IV: Thực Nghiệm Và Kết Quả 20**](#_9kzdhcxh227g)

[**V: Kết Luận Và Hướng Phát Triển 20**](#_a1bkeabbfybd)

**LỜI CẢM ƠN**

Nhóm chúng em xin chân thành cảm ơn **Trường Đại học Sài Gòn** đã tạo điều kiện thuận lợi về cơ sở vật chất và môi trường học tập để chúng em có thể thực hiện và hoàn thành bài báo cáo này.

Đặc biệt, nhóm xin gửi lời cảm ơn sâu sắc đến **TS.Đỗ Như Tài** – giảng viên hướng dẫn, người đã tận tình chỉ bảo, định hướng và hỗ trợ chúng em trong suốt quá trình thực hiện đề tài. Những ý kiến đóng góp quý báu của Thầy là kim chỉ nam giúp nhóm hoàn thiện bài báo cáo một cách tốt nhất.

Mặc dù nhóm đã nỗ lực hết mình, nhưng do hạn chế về kiến thức và kinh nghiệm, bài báo cáo chắc chắn vẫn còn những thiếu sót. Nhóm rất mong nhận được sự góp ý từ quý Thầy Cô để hoàn thiện hơn trong những lần sau.

**Nhóm chúng em xin chân thành cảm ơn!**

# 

# **Chương I: Tổng Quan Vấn Đề**

1. **Lý do chọn đề tài**

Trong thời đại số hóa hiện nay, các nền tảng thương mại điện tử đang ngày một phát triển và nắm giữ một vai trò quan trọng trong việc kết nối các doanh nghiệp với khách hàng. Một trong những yếu tố ảnh hưởng trực tiếp đến quyết định mua hàng của khách hàng đó chính là những đánh giá, nhận xét từ những khách hàng trước đó.

Tuy nhiên, số lượng đánh giá được tạo ra mỗi ngày vô cùng lớn khiến việc phân tích thủ công gặp nhiều khó khăn. Do đó việc phân tích cảm xúc (Sentiment Analysis) kết hợp với các phương pháp học máy (Machine Learning), đã trở thành một công cụ mạnh mẽ giúp tự động hóa quá trình phân tích và phân loại cảm xúc của khác hàng. Việc áp dụng học máy không chỉ giúp các doanh nghiệp hiểu rõ hơn về mức độ hài lòng của khách mà còn hỗ trợ việc cải tiến sản phẩm, tạo chiến lược phát triển phù hợp.

1. **Vấn đề nghiên cứu**

Việc phân tích cảm xúc của người dùng hiện nay gặp phải nhiều thách thức. Một trong những vấn đề lớn là các mô hình học máy khác nhau sẽ cho ra những kết quả khác nhau, và việc lựa chọn mô hình phù hợp nhất để đạt được hiệu quả là vô cùng quan trọng. Bên cạnh đó, quá trình tiền xử lý văn bản cũng đóng vai trò then chốt trong việc nâng cao chất lượng phân tích. Một bài đánh giá có thể chứa rất nhiều thông tin, trong đó một số thông tin có thể không liên quan đến sản phẩm, gây nhiễu cho quá trình phân tích.

1. **Mục tiêu nghiên cứu**

Trước thực trạng được nêu ra, mục tiêu của nhóm nghiên cứu là triển khai và đánh giá các mô hình học máy nhằm so sánh hiệu suất và độ chính xác giữa các phương pháp khác nhau. Cụ thể, nhóm sẽ xây dựng và áp dụng một số mô hình học máy như **Logistic Regression** và **Naive Bayes** trên tập dữ liệu đánh giá sản phẩm thực phẩm từ Amazon.

Các mô hình sẽ được đánh giá dựa trên các chỉ số như **Accuracy**, **Precision**,... nhằm xác định mô hình tối ưu nhất. Từ đó, nhóm đề xuất một mô hình phù hợp có thể tích hợp vào các hệ thống thực tế, hỗ trợ doanh nghiệp trong việc phân tích cảm xúc khách hàng, cải thiện sản phẩm và nâng cao trải nghiệm người dùng.

1. **Đối tượng và phạm vi nghiên cứu**

Trong bài nghiên cứu này nhóm sẽ tập trung vào việc phân tích các đánh giá có trong tập dữ liệu Amazon Fine Food Reviews - một tập đánh giá uy tín , chứa hàng ngàn đánh giá thực phẩm của người dùng.

Nhóm sẽ áp dụng các kĩ thuật xử lí ngôn ngữ tự nhiên (NLP) để khai phá các dữ liệu từ đó so sánh hiệu quả của các mô hình học máy. Đồng thời nhóm cũng áp dụng kĩ thuật trích xuất đặc trưng văn bản để xác định các thông tin quan trọng trong việc đánh giá, góp phần nâng cao độ chính xác cho quá trình phân loại.

Nghiên cứu của nhóm chỉ tập trung vào việc phân loại cảm xúc của dữ liệu đó là tích cực, tiêu cực, trung lập.

1. **Phương pháp nghiên cứu**

Bài nghiên cứu này là sự kết hợp chặt chẽ giứa các phương pháp lý thuyết , phương pháp thực nghiệm, phương pháp chuyên gia nhằm đảm bảo tính khoa học , tính nhất quán và hệ thống trong quá trình nghiên cứu đề tài.

**5.1. Phương pháp lý thuyết**

Tìm hiểu các bài báo có liên quan đền chủ đề phân tích cảm xúc bằng mô hình máy học thông qua các bài nghiên cứu, bài báo trên các nguồn uy tín.

Tổng hợp và phân tích các cơ sở lý thuyết về kĩ thuật xử lí ngôn ngữ tự nhiên (NLP), trích xuất đặc trưng và các khái niêm các mô hình máy học .

Tìm đến các nguồn báo uy tín từ các tạp chí khoa học hàng đầu như IEEE, Springer và các hội nghị trí tuệ nhân tạo (NeurIPS, ACL, EMNLO) để đảm bảo sự uy tín cho bài nghiên cứu.

**5.2. Phương pháp thực nghiệm**

Phân tích , thiết kế triển kahi các mô hình học máy trên tập dữ liệu Amazon Fine Food Reviews

Cài đặt các mô hình và thuật toán bằng python và sử dụng các thư viện liên quan

Thực hiện huấn luyện và kiểm thử các mô hình, ghi nhận và đánh giá hiệu suất các mô hình bằng độ đo

Thực hiện phân tích so sánh các mô hình và lựa chọn mô hình phù hợp với thực tế nhất

**5.3. Phương pháp chuyên gia**

Trao đổi với giảng viên hướng dẫn để làm rõ các khuất mắt trong qui trình thực hiện đồng thời tiếp nhận các đánh giá và phản hổi nhằm cải thiện bài nghiên cứu và đưa ra mục tiêu , nhiêm vụ phù hợp cho từng giai đoạn

1. **Kế hoạch nghiên cứu dự kiến**
2. **Cấu trúc bài nghiên cứu**

# **Chương II: Lược khảo tài liệu**

1. **Cơ sở lý thuyết**

**1.1. Tổng quan về phân tích cảm xúc**

**Khái niệm phân tích cảm xúc**

Phân tích cảm xúc (sentiment analysis) là một nhánh quan trọng của xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc trích xuất và phân loại các ý kiến, cảm xúc hoặc thái độ được thể hiện trong văn bản.

Để thực hiện, quá trình này thường áp dụng các phương pháp dựa trên từ điển, học máy hoặc học sâu, cho phép xử lý hiệu quả khối lượng lớn dữ liệu văn bản từ các nguồn như mạng xã hội hoặc đánh giá sản phẩm.

**Ứng dụng của phân tích cảm xúc**

Phân tích cảm xúc đóng vai trò quan trọng trong kinh doanh, đặc biệt là trong lĩnh vực thương mại điện tử, được ghi nhận qua nhiều nghiên cứu . Các doanh nghiệp sử dụng phân tích cảm xúc để đánh giá phản hồi khách hàng từ đánh giá sản phẩm trên các nền tảng như Amazon hoặc Shopee, từ đó cải thiện chất lượng sản phẩm và dịch vụ. Ngoài ra, công nghệ này hỗ trợ theo dõi danh tiếng thương hiệu trên mạng xã hội, tối ưu hóa trải nghiệm khách hàng và dự đoán xu hướng thị trường thông qua phân tích hành vi người tiêu dùng trực tuyến . Những ứng dụng này giúp doanh nghiệp nâng cao lợi thế cạnh tranh và đáp ứng tốt hơn nhu cầu của thị trường thương mại điện tử.

**1.2. Các cấp độ của việc phân tích cảm xúc**

Các cấp độ của việc phân tích cảm xúc được chia thành 3 cấp độ:

**Cấp độ tài liệu (Document-level Sentiment Analysis):** Ở cấp độ này, toàn bộ văn bản được xem như một đơn vị duy nhất. Mục tiêu của cấp độ này là xác định tổng thể văn bản thể hiện cảm xúc tích cực, tiêu cực, trung lập đối với một đối tượng. Cấp độ này sẽ phù hợp nếu văn bản chỉ thể hiện một quan điểm duy nhất về một đối tượng.Theo Pang và Lee (2004), phương pháp này phù hợp với các hệ thống đánh giá đơn giản, nhưng phương pháp này sẽ không hiệu quả nếu văn bản chứa nhiều quan điểm khác nhau.

**Cấp độ sâu (Sentence-level Sentiment Analysis):** Ở cấp độ này, các câu trong văn bản được tách riêng để xác định cảm xúc mà câu đó diễn đạt. Phương pháp này phù hợp với văn bản có nhiều câu và các câu đó thể hiện các quan điểm trái ngược nhau. Theo Kim và Hovy (2004) , việc phân tích ở cấp độ câu giúp tăng độ chính xác khi phân tích các văn bản dài, nhưng phương pháp này chưa xử lý tốt mối quan hệ giữa cảm xúc và thực thể.

**Cấp độ đặc trưng (Aspect-level Sentiment Analysis):** Đây là cấp độ được phân tích chi tiết nhất , không chỉ xác định cảm xúc hệ thống còn liên kết cảm xúc đó với khía cạnh cụ thể của sản phẩm. Phương pháp này phù hợp với các hệ thống đánh giá sản phẩm có nhiều thuộc tính và đặc trưng khác nhau.

**1.3. Phương pháp phân tích cảm xúc**

Phương pháp phân tích cảm xúc được chia là 3 nhóm đó chính là

**Mô hình dựa trên quy tắc (Rule-based / Lexicon-based Methods):** Phương pháp sử dụng từ điển cảm xúc (sentiment lexicons) chứa các từ được gắn sẵn điểm số cảm xúc. Hệ thống sẽ tính tổng điểm cảm xúc trong văn bản dựa trên sự xuất hiện và trọng số của các từ trong từ điểm. Đây là một phương pháp rất dễ xây dựng , không cần huấn luyện nhiều nhưng tỏ ra không linh hoạt khi xử lí các văn bản chứa nghĩa bóng, ẩn ý và có ngữ cảnh phức tạp.

**Mô hình học máy (Machine Learning-based Methods):** Dùng các thuật toán Machine Learning để học các đặc trưng từ dữ liệu văn bản đã gắn nhãn. Mô hình này có độ chính xác cao hơn mô hình dựa trên quy tắc nhưng đòi hỏi tập dữ liệu lớn và bị phụ thuộc vào đặc trưng đầu vào (feature engineering).

**Mô hình học sâu (Deep Learning-based Methods):** Ứng dụng các mô hình mạng neuron sâu (deep neural networks) để tự động học biểu diễn ngữ nghĩa và cấu trúc ngôn ngữ từ dữ liệu. Mô hình này có khả năng tự động trích xuất đặc trưng đặc trưng, xử lý tốt các văn bản có ngữ cảnh phức tạp nhưng đòi hỏi tài nguyên tính toán lớn và dữ liệu huấn luyện khổng lồ.

**1.4. Kỹ thuật trích xuất đặc trưng**

**Bag of Words (BoW):** Là một kỹ thuật trong xử lý ngôn ngữ tự nhiên (NLP) để biểu diễn văn bản dưới dạng một tập hợp các từ mà không quan tâm đến ngữ pháp hoặc vị trí của từ trong câu.

Ta có thể giải thích cách hoạt động của kỹ thuật này qua 2 câu ví dụ “I love this movie” và “This movie is amazing”:

Bước 1: Chuỗi văn bản sẽ được phân tách thành các từ riêng lẻ, sau đó tạo thành một danh sách chứa tất cả các từ xuất hiện trong văn bản

["I", "love", "this", "movie", "is", "amazing"]

Bước 2: Biểu diễn văn bản dưới dạng vector. Chuyển mỗi câu thành một vector đếm số lần xuất hiện của từ vựng trong câu đó.

|  | I | love | this | movie | is | amazing |
| --- | --- | --- | --- | --- | --- | --- |
| I love this movie | 1 | 1 | 1 | 1 | 0 | 0 |
| This movie is amazing | 0 | 0 | 1 | 1 | 1 | 1 |

**1.5. Khái niệm các mô hình học máy**

**Logistic Regression**:

**Naive Bayes**:

**1.6. Độ đo**

**Khái niệm độ đo**

Độ đo (metric) là các tiêu chí dùng để đánh giá hiệu suất của các mô hình.

**Phân loại độ đo**

Các độ đo được áp dụng tùy thuộc vào loại bài toán được phân làm 3 nhóm chính:

**Độ đo trong bài toán phân loại**

**Accuracy (Độ chính xác)**:

**Precision (Độ chính xác lớp dương)**:

**Recall (Độ nhạy)**:

**F1-score**:

**2. Tổng hợp và đánh giá các bài nghiên cứu trước đó**

Phân tích cảm xúc (Sentiment Analysis) trong lĩnh vực thương mại điện tử đã và đang thu hút sự quan tâm từ các cộng đồng khoa học. Với số lượng đánh giá sản phẩm trực tuyến tăng cao, việc xử lý một khối lượng lớn các đánh giá theo thu công là bất khả thi. Trong những năm gần đây, nhiều công trình đã được công bố, đem lại sự đóng góp lớn trong việc cải thiện độ chính xác và hiệu quả cao, trở thành giải pháp quan trọng trong việc tự động phân loại đánh giá của người dùng từ việc sử dụng mô hình Machine Learning và Deep Learning trong phân loại cảm xúc người dùng từ các đánh giá sản phẩm.

Ở nước ngoài, các nghiên cứu tập trung vào việc cải thiện độ chính xác và hiệu quả của các mô hình phân tích cảm xúc. Nghiên cứu của Zhao và Sun đã sử dụng BERT (Bidirectional Encoder Representations from Transformers) với fine-tuning BERT để phân tích cảm xúc từ các đánh giá sản phẩm trực tuyến [4]. Nghiên cứu cho thấy việc tận dụng khả năng trích xuất đặc trưng có thể cải thiện độ chính xác cao trong việc phân loại cảm xúc. Với hướng tiếp cận khác của Aravindan và nhóm tác giả của mình đã sử dụng PySpark, một công cụ xử lý dữ liệu lớn, để phân loại cảm xúc đánh giá người dùng [5]. Nghiên cứu này cho thấy việc sử dụng các phương pháp phân tích dữ liệu đem loại hiệu quả tốt nhưng vẫn còn tồn tại nhiều thách thức về hiệu suất và tối ưu. Các nghiên cứu quốc tế ứng dụng ML và NLP để phân tích cảm xúc, ứng dụng các phương pháp học sâu và xử lý dữ liệu lớn. Tuy nhiên các thách thức về độ chính xác, khả năng xử lý dữ liệu vẫn còn khá nhiều hạn chế.

Trong nước, lĩnh vực phân tích cảm xúc đang thu hút sự chú ý của các nhà nghiên cứu, đặc biệt trong bối cảnh thương mại điện tử đang phát triển mạnh mẽ. Các công trình nghiên cứu có sử dụng các mô hình học máy để phân loại dữ liệu để ứng dụng NLP vào Tiếng Việt [6][7]. Đáng chú ý với nghiên cứu của Trần Quang Phúc và nhóm của mình đã cho thấy việc tập trung trích xuất các từ ngữ (Opinion words) trong các đánh giá giúp nâng cao hiệu quả phân tích cảm xúc [8]. Nghiên cứu sử dụng các thuật toán học máy theo phương pháp tập hợp (ensemble methods) như Gradient Boosting Classifier (GBC) đem lại độ chính xác cao trong giai đoạn kiểm tra, giúp cải thiện khả năng đánh giá cảm xúc tích cực và tiêu cực. Lĩnh vực phân tích cảm xúc đang có những bước tiến đáng kể nhưng vẫn còn rất nhiều tìêm năng để phát triển, mở ra các cơ hội khai thác dữ liệu tốt hơn.

# **III: Phương Pháp Nghiên Cứu**

1. **Quy trình nghiên cứu và các bước thực hiện**

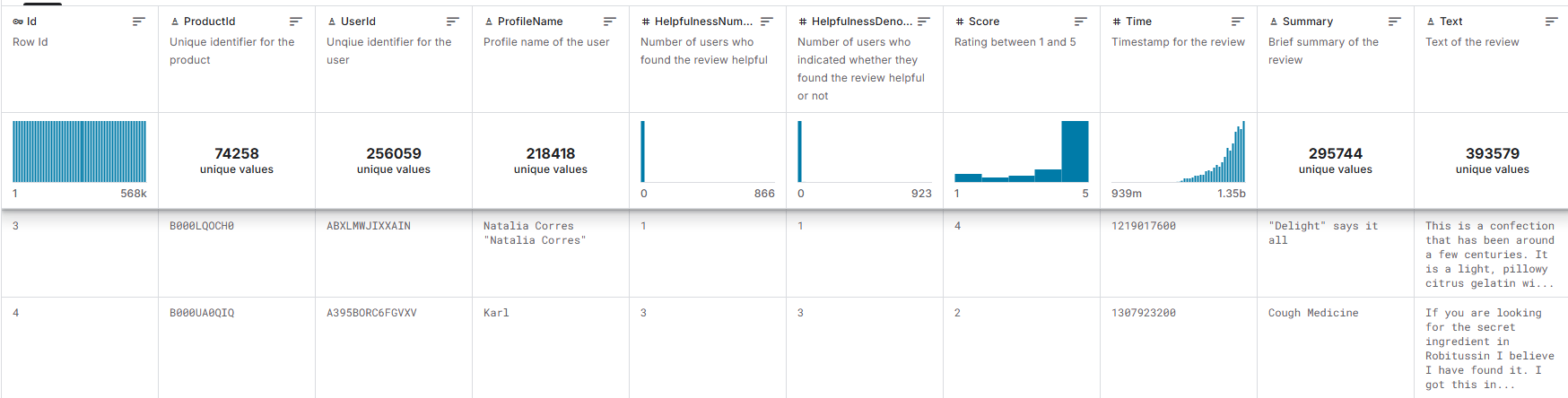
Qui trình nghiên cứu bao gồm các bước sau:

* Khảo sát dữ liệu
* Tiền xử lí dữ liệu
* Trích xuất đặc trưng
* Huấn luyện mô hình
* Phân tích và so sánh các mô hình
* Chọn ra mô hình phù hợp nhất

Sơ đồ thể hiện qui trình nghiên cứu

1. **Nguồn dữ liệu**

Nghiên cứu này sử dụng bộ dữ liệu Amazon Fine Food Reviews thu thập từ Kaggle, bao gồm các đánh giá về sản phẩm thực phẩm trên nền tảng Amazon. Bộ dữ liệu này trải dài trong hơn 10 năm, từ năm 2002 đến tháng 10 năm 2012, với tổng cộng hơn 500.000 đánh giá. Mỗi đánh giá cung cấp các thông tin quan trọng như thông tin sản phẩm (tên sản phẩm, danh mục), thông tin người dùng (ID người dùng), xếp hạng (từ 1 đến 5), và nội dung đánh giá dạng văn bản.

****

Tập dữ liệu AFFR có 10 cột dữ liệu, nội dung các cột dữ liệu là:

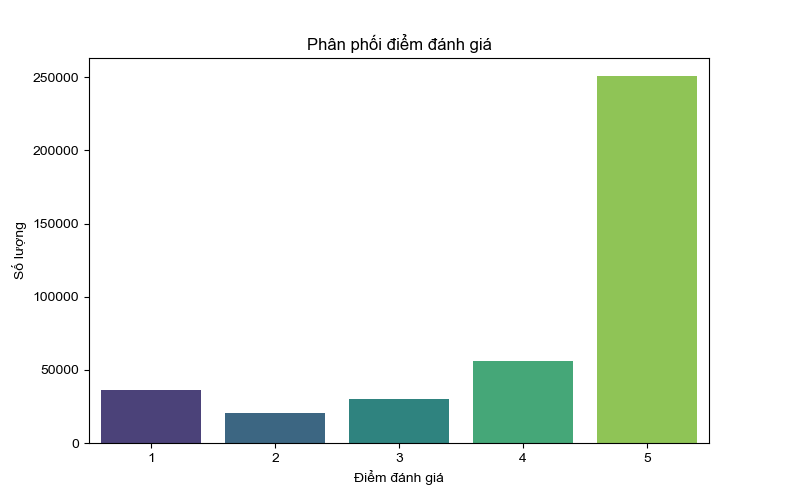
| Id | Mã Id định danh cho mỗi đánh giá |
| --- | --- |
| ProductId | Id sản phẩm cho mỗi đánh giá |
| UserId | Id người dùng đã đánh giá |
| ProfileName | Tên người đánh giá |
| HelpfulnessNumerator | Số lượng người thấy đánh giá hữu ích |
| HelpfulnessDenominator | Tổng số lượt người dùng đã bỏ phiếu mức độ hữu |
| Score | Điểm đánh giá từ 1 đến 5 |
| Time | Mốc thời gian đánh giá theo dạng Unix time |
| Summary | Phần tóm tắt ngắn hoặc tiêu đề của đánh giá |
| Text | Nội dung đầy đủ của đánh giá |

1. **Khảo sát dữ liệu**

Giai đoạn khảo sát dữ liệu đóng vai trò làm nền tảng trong quá trình phân tích. Giai đoạn này báo gồm các phân tích cơ bản như:

**3.1. Univariate analysis (phân tích đơn biến)**

Mục tiêu của phân tích này đó là kiểm tra phân phối của điểm đánh giá (Score) theo mức độ từ 1 đến 5 sao



Nhận xét:

* Điểm số đánh giá thiên về đánh giá 5 sao
* Các đánh giá 1 sao, 2 sao chiếm tỉ lệ thấp
* Khoảng hơn 250 000 đánh giá tốt, cho thấy khách hàng có trải nghiệm tốt với sản phẩm
* Tập dữ liệu có thể bị mất cân bằng khi thực hiện các mô hình học máy, cần xem xét việc cân bằng dữ liệu (resampling)

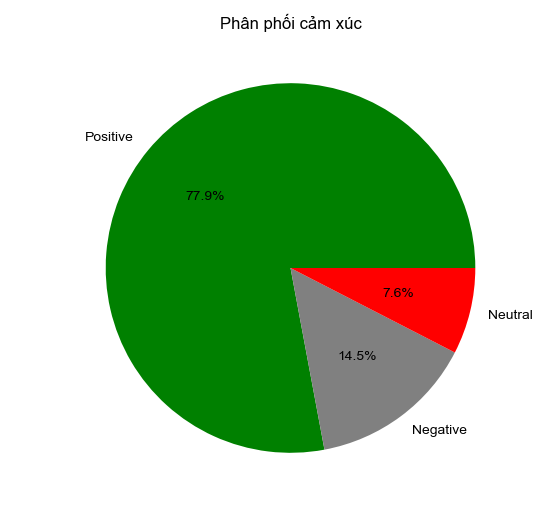
**3.2. Sentiment distribution analysis (phân tích phân phối cảm xúc)**

Mục tiêu của phân tích là chuyển đổi điểm số thành các nhãn tích cực (positive), tiêu cực (negative) và trung lập (neutral)

Positive: Điểm đánh giá >3

Neutral: Điểm đánh giá =3

Negative: Điểm đánh giá <3



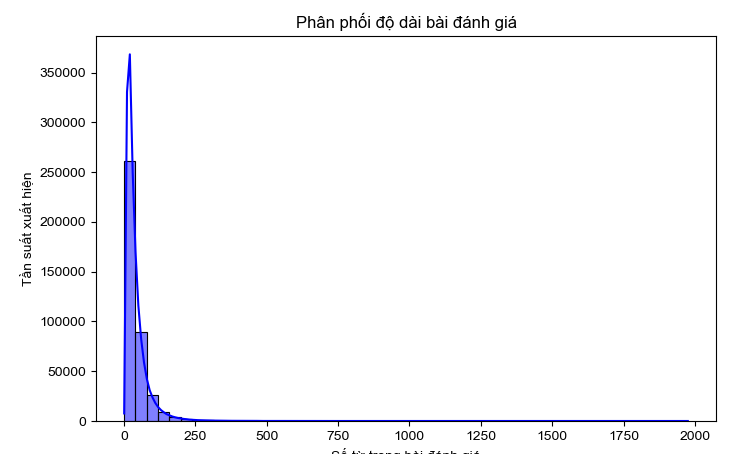
Nhận xét:

* Đánh giá tích cực (4,5) chiếm đa số với khoảng 77.9%
* Đánh giá tiêu cực (1,2) chỉ chiếm khoảng 14.5%
* Tỷ lệ đánh giá trung lập (3 sao) chỉ chiếm 7.6%, cho thấy phần lớn khách hàng có ý kiến rõ ràng thay vì đánh giá trung tính.
* Tập dữ liệu lệch về hướng tích cực, có thể gây ảnh hưởng đến mô hình dễ dự đoán tích cực hơn tiêu cực

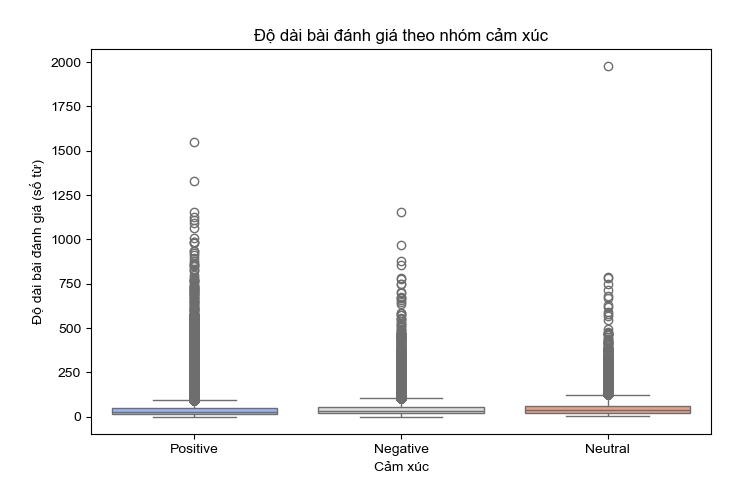
# **3.3. Review length analysis (độ dài các đánh giá)**

Mục tiêu là kiểm tra mối quan hệ giữa độ dài đánh giá (Text) và điểm số (Score) của người dùng.

Biểu đồ Histogram độ dài đánh giá (tượng trưng cho độ dài số từ)



Biểu đồ Boxplot so sánh độ dài giữa các nhóm đánh giá điểm số



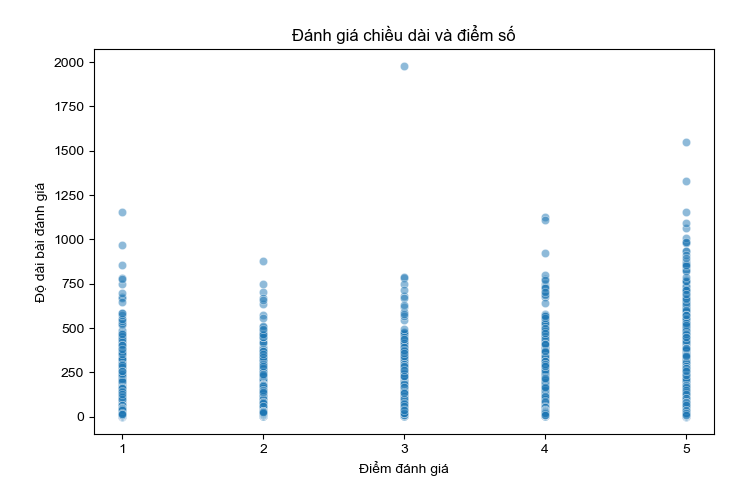
Nhận xét:

* Từ biểu đồ đầu tiên ta thấy độ dài đánh giá tập trung vào 0-100 số từ.
* Một số bài đánh giá có độ dài rất lớn, có thể do spam
* Khi so sánh với nhóm đánh giá dựa trên các nhãn, các đánh giá tích cực có độ dài trung bình cao hơn tiêu cực
* Các đánh giá tiêu cực ngắn hơn (khoảng 30 ~ 50 từ)
* Một số đánh giá có dài rất cao, có thể là người dùng spam
* Người dùng có xu hướng viết đánh giá dài khi họ cảm thấy hài lòng với thực phẩm.
* Trước khi huấn luyện mô hình, có thể cần lọc các bài quá dài để không ảnh hưởng đến kết quả dự đoán

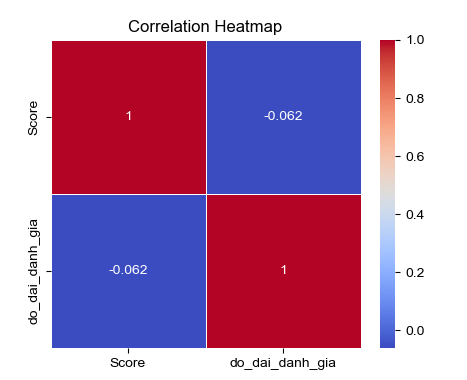
# **3.4. Correlation analysis (phân tích tương quan)**

Mục tiêu khi phấn tích là kiểm tra các yếu tố có ảnh hưởng đến điểm số đánh giá của người dùng

Biều đồ Scatter plot so sánh độ dài đánh giá(Text) và điểm số(Score)



Biểu đồ Heatmap ma trận tương quan giữa độ dài đánh giá(Text) và điểm số(Score)



Nhận xét:

* Biểu đồ phân tán các điểm dữ liệu phân bố rời rạc theo các mức điểm
* Không có xu hướng thể hiện mối quan hệ giữa độ dài bài đánh giá và điểm số.
* Biểu đồ Heatmap với hệ số tượng quan âm (-0.062) cho thấy không có mối quan hệ quá mạnh giữa độ dài đánh giá và điểm số.
* Dữ liệu không thể hiện các bài đánh giá dài hơn thì điểm đánh giá sẽ cao hơn.

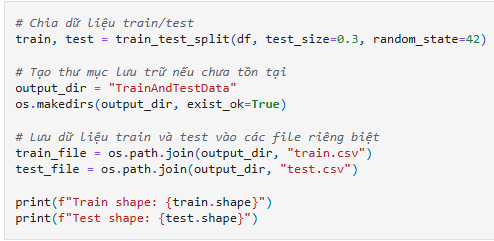
# **3.5. Conclusion (kết luận)**

Sau khi thực hiện các bước phân tích khai phá dữ liệu ta có các ý sau:

* Dữ liệu chứa các đánh giá sản phẩm với các thông tin quan trọng như Nội dung đánh giá (Text), Điểm đánh giá (Score), và Cảm xúc trong đánh giá (Sentiment).
* Dữ liệu đánh giá có sự mất cân bằng: phần lớn điểm số ở mức 5 sao, dẫn đến nhiều đánh giá có cảm xúc tích cực hơn tiêu cực.
* Điểm đánh giá 3 (Neutral) không đóng góp trong việc phân tích, có thể gây cản trở huấn luyện mô hình  
  Số lượng từ trong đánh giá ngắn chiếm đa số (dưới 100 từ), tuy nhiên các bài đánh giá tích cực và tiêu cực có xu hướng dài hơn đánh giá trung lập.
* Không có mối tương quan rõ ràng giữa điểm số và độ dài bài đánh giá (hệ số tương quan -0.062), nghĩa là độ dài không quyết định mức điểm mà người dùng đánh giá.
* Có thể cần cân bằng lại dữ liệu hoặc áp dụng các phương pháp xử lý đặc biệt để đảm bảo các mô hình phân tích dự đoán chính xác hơn.

**4. Tiền xử lí dữ liệu và trích xuất đặc trưng**

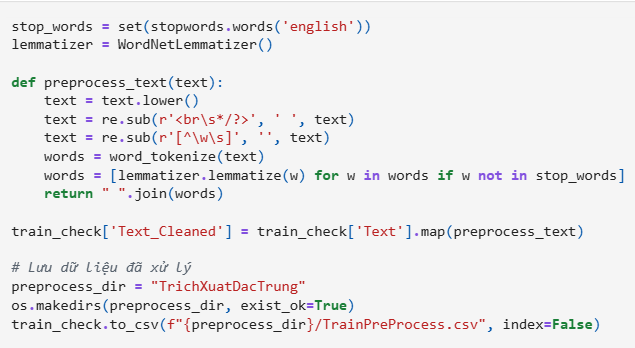
Trước khi tiền xử lí và trích xuất đặc trưng nhóm quyết định sẽ chia nguồn dữ liệu thành 2 tập train và test



**4.1. Tiền xử lí dữ liệu**

Quá trình tiền xử lí dữ liệu vô cùng quan trọng giúp loại bỏ những dữ liệu có thể gây ảnh hưởng đến quá trình phân tích của các mô hình quá trình bao gồm các bước đó là:

* Chuyển văn bản thành chữ thường
* Xóa thẻ HTML
* Xóa dấu câu
* Tách từ
* Loại bỏ stopwords
* Lemmatization (chuyển từ về dạng gốc)

****

Ví dụ trước và sau khi tiền xử lí:

* Trước tiền xử lí

I found this in a store but then they discontinued it where I am. It is a great alternative to hot chocolate or coffee. Not too sweet and has great flavor.

* Sau khi tiền xử lí

found store discontinued great alternative hot chocolate coffee sweet great flavor

**4.2. Trích xuất đặc trưng**

Sau khi tiền xử li đến quá trình trích xuất đặc trưng bao gồm 3 bước:

* Áp dụng CountVectorizer để chuyển đổi văn bản đã tiền xử lí thành ma trận đặc trưng Bag of Words (BoW) với tối đa 1000 từ phổ biến nhất.

Mỗi cột trong ma trận đặc trưng sẽ ứng với một từ vựng được trích ra từ tập dữ liệu và giá trị mỗi ô sẽ biểu diễn tần xuất xuất hiện của từ đó trong văn bản.

* Bổ sung thêm 2 đặc trưng nữa là

Text\_Length: Tổng số kí tự trong văn bản gốc Text

Word\_Count: Tổng số từ trong văn bản đã được làm sạch Text\_Cleaned

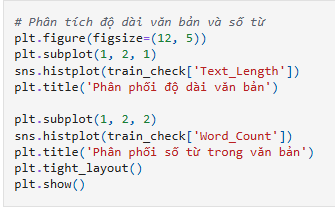
- Kết hợp và lưu trữ các đặc trưng



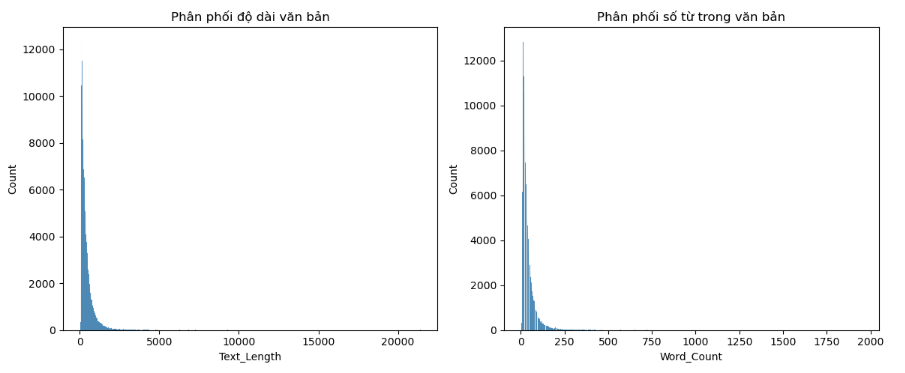
**4.3. Phân tích các đặc trưng**

**Phân tích sự phân phối độ dài văn bản và số từ**

Mục đich của phân tích này đó là xác định xem dữ liệu có nhiều văn bản cực ngắn hoặc cực dài hay không đồng thời hiểu được mức độ đa dạng của văn bản



Sơ đồ

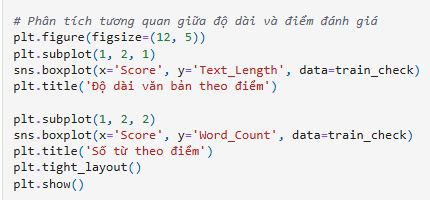


Nhận xét

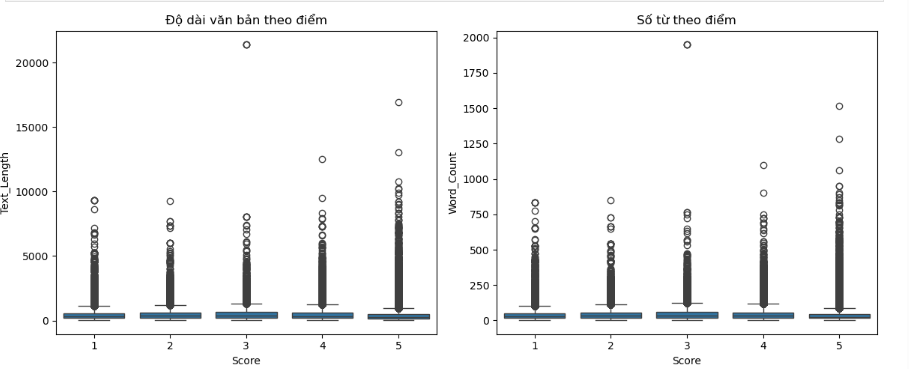
* Cả hai sơ đồ đều bị lệch phải (Nghĩa là có rất nhiều văn bản văn bản ngắn và ít văn bản dài)
* Một số văn bản có độ dài hoặc số từ rất cao, cần xem xét loại bỏ hoặc chuẩn hóa.

**Phân tích mối tương quan giữa độ dài/ số từ với điểm đánh giá**

Mục tiêu là xác định xem liệu văn bản hoặc số từ có liên quan đền điểm đánh giá hay không



Sơ đồ



Nhận xét:

* Các boxplot cho thấy độ dài văn bản và số từ **không thay đổi quá nhiều** giữa các mức điểm 1 đến 5.

**5. Huấn luyện mô hình**

# **IV: Thực Nghiệm Và Kết Quả**

1. **Trình bày kết quả phân tích thu được**
2. **Đánh giá và giải thích kết quả**
3. **Ý nghĩa thực tiễn của nghiên cứu**
4. **Chỉ ra các vấn đề hạn chế**

# **V: Kết Luận Và Hướng Phát Triển**

1. **Những điểm đã đạt được**
2. **Đề xuất hướng nghiên cứu tiếp theo**

Danh Mục Các Từ Viết Tắt

Danh Mục Hình Ảnh

Danh Mục Các Tài Liệu Tham Khảo