ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN

PHÂN LOẠI ĐA NHÃN ĐÁNH GIÁ CỦA KHÁCH HÀNG TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ TIKI

Sinh viên thực hiện:

Nguyễn Phan Trúc Quỳnh – 21522536

Nguyễn Minh Quân - 21522496

Nguyễn Thị Cẩm Ly – 21522315

Lương Triệu Hoàng Vũ - 21522798

Giảng viên:

ThS. Nguyễn Văn Kiệt

Thành phố Hồ Chí Minh, tháng 12 năm 2023

MỤC LỤC

TÓM LƯỢC

Trong báo cáo này, chúng tôi tập trung phân loại đa nhãn đánh giá của khách hàng trên sàn thương mại điện tử Tiki. Việc phân loại đa nhãn đánh giá của khách hàng nhằm mục đích xác định những trải nghiệm, suy nghĩ khác nhau của khách hàng về sản phẩm mà họ đã mua và sử dụng. Đầu tiên, chúng tôi đã xây dựng bộ dữ liệu từ gần 3500 đánh giá của khách hàng về sản phẩm thuộc lĩnh vực chăm sóc da và trang điểm trên trang web Tiki. Sau đó, tiến hành huấn luyện mô hình phân loại đa nhãn bằng các thuật toán Naïve Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbors, Logistic Regression. Cuối cùng, chúng tôi so sánh và phân tích kết quả thu được bằng cách sử dụng một bộ số liệu thống kê kết quả. Theo kết quả nghiên cứu thử nghiệm, chúng tôi nhận thấy phương pháp là Binary Relevance sử dụng thuật toán RF cho kết quả Micro F1 Score 0,85914 và Hamming Loss 0.05078 là phương pháp mang lại kết quả tốt nhất.

CHƯƠNG 1: TỔNG QUAN

1. Tổng quan về đề tài

Nhiệm vụ phân loại đa nhãn đánh giá của khách hàng nhằm mục đích xác định những trải nghiệm, suy nghĩ khác nhau của khách hàng về sản phẩm mà họ đã mua và sử dụng. Trong thời đại kỹ thuật số hiện nay, việc sử dụng trang web để mua sắm đã và đang rất phổ biến, xu hướng mua hàng tại các trang thương mại điện tử tăng mạnh. Dẫn đến lượng dữ liệu khách hàng trên các trang này không ngừng tăng giúp cho việc thực hiện các nghiên cứu.

Các tiến bộ trong NLP và công nghệ thông tin đã tạo điều kiện cho sự ra đời của các ứng dụng mới, đồng thời mở rộng không gian nghiên cứu dựa trên dữ liệu văn bản ngày càng phong phú. Phân loại văn bản là một trong những bài toán cơ bản của NLP. Phân loại văn bản là việc gán các nhãn đã có sẵn cho các đoạn văn bản tương ứng. Tùy thuộc vào từng trường hợp, mỗi văn bản có thể được gắn một hoặc nhiều nhãn. Sự đa dạng các cách tiếp cận dẫn đến các kiểu phân loại văn bản khác nhau.

Có nhiều cách phân loại văn bản trong NLP, nếu dựa trên số lượng nhãn được gán cho mỗi văn bản, có thể chia phân loại văn bản thành 2 loại là phân loại nhị phân (Binary) hay phân loại đa lớp (Multi-class) và phân loại đa nhãn (Multi-label). Khác với phân loại nhị phân hay phân loại đa lớp, mỗi văn bản chỉ được gán một nhãn, thì với phân loại đa nhãn, mỗi văn bản có thể được gán nhiều nhãn khác nhau.

Kiểu phân loại	Bài toán	Nhãn
Phân loại nhị phân	Phân loại tin thật, giả	Thật, giả
Phân loại đa lớp	Phân tích cảm xúc	Tích cực, tiêu cực, trung lập
Phân loại đa nhãn	Phân loại cảm xúc	Buồn, vui, giận, sợ hãi

Bảng 1: Ví dụ về các bài toán phân loại văn bản.

Đối với ngữ cảnh cần thu thập ý kiến người dùng về một sản phẩm hay một vấn đề, bài toán phân loại nhị phân hay phân loại đa lớp sẽ tỏ ra kém hiệu quả vì trong những trường hợp này ngôn ngữ mà chúng ta sử dụng phức tạp hơn rất nhiều, gắn nhãn đơn

lẻ cho văn bản sẽ làm hạn chế khả năng trích xuất thông tin. Vì vậy sử dụng bài toán phân loại đa nhãn trở nên phù hợp.

Theo thống kê từ [2], ước tính đến năm 2024, 95% giao dịch sẽ diễn ra trên các trang thương mại điện tử. Đánh giá sản phẩm có ảnh hưởng lớn đến quyết định mua sắm của khách hàng [3]. 90% khách hàng đọc đánh giá trực tuyến trước khi mua hàng. Lượng dữ liệu khổng lồ và tác động của nó đến khách hàng là nguyên nhân dẫn đến nhiều nghiên cứu về phân loại đánh giá khách hàng thương mại điện tử được thực hiện. Phần lớn các nghiên cứu trước đây trong lĩnh vực này chỉ phân tích độ tích cực (positive hay negative) của đánh giá, chưa thực hiện phân loại đa nhãn.

Ở bài báo cáo này, chúng tôi đề xuất phương pháp để xác định các nhãn khác nhau có trong đánh giá của người dùng. Phân loại đa nhãn giúp phân tích đánh giá khách hàng thương mại điện tử một cách chi tiết và sâu sắc hơn, mang lại lợi ích cho cả người mua hàng và doanh nghiệp. Bộ dữ liệu mới và các phương pháp phân tích mà bài báo cáo sử dụng có thể được áp dụng cho các nghiên cứu tương lai trong lĩnh vực này.

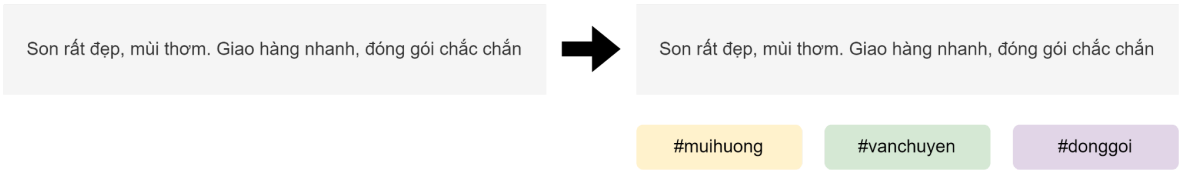
2. Mô tả bài toán

Đề tài phân loại đa nhãn đánh giá của khách hàng trên sàn thương mại điện tử tiki hướng đến việc, từ bình luận của người mua hàng, có thể tìm ra được những thông tin cụ thể mà khách hàng đề cập trong bình luận, từ đó có thể phân loại đánh giá của người dùng. Thay vì chỉ phân loại đánh giá tích cực hay tiêu cực, chúng tôi muốn đi sâu vào những thông tin mà khách hàng đã đề cập để hiểu chi tiết hơn quan điểm của khách hàng. Toàn bộ dữ liệu được chúng tôi thu thập từ trang thương mại điện tử tiki, nhiệm vụ của bài toán là phân loại đánh giá của khách hàng thành 8 nhãn (cụ thể sẽ được trình bình bên dưới), với mỗi đánh giá có thể thuộc một hay nhiều nhãn. Đề tài này có thể làm nền tảng cho việc xây dựng bộ lọc đánh giá của khách hàng. Trong ngữ cảnh Machine Learning, đây là dạng bài toán Multi-Label Classification, với 8 nhãn khác nhau và 2 classes là 0 và 1.

3. Input và output của bài toán

Input: đánh giá của khách hàng trên sàn thương mại điện tử Tiki

Output: đánh giá từ input kèm các nhãn liên quan.



4. Các thuật toán máy học mà đồ án sử dụng

- Naïve Bayes (NB):

  NB là một thuật toán máy học được dựa trên định lý Bayes về giả định tính độc lập giữa các đặc trưng. Thuật toán này dựa trên xác suất để dự đoán hay phân lớp nhãn của một mẫu dựa trên xác suất tiên nghiệm và xác suất hậu nghiệm. NB có cơ chế hoạt động dựa trên định lý Bayes, một định lý cơ bản trong xác suất thống kê. Công thức Bayes cho phép tính xác suất hậu nghiệm (xác suất của một biến cố xảy ra sau khi có thông tin mới) dựa trên xác suất tiên nghiệm (xác suất của một biến cố diễn ra trước khi có thông tin mới) và thông tin mới đó. Xác suất hậu nghiệm được tính như sau:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

  Trong đó, P(y|X) là xác suất của lớp y khi biết X, P(X|y) là xác suất của lớp X khi biết y, P(y) là xác suất tiên nghiệm của lớp y và P(X) là xác suất đặc trưng X. Thông qua việc tính toán xác suất diễn ra của từng lớp, ta có thể tiến hành phân loại dựa vào phân lớp có xác suất cao nhất.

- Support Vector Machine (SVM):

  SVM là một mô hình phân loại hoạt động bằng việc xây dựng một siêu phẳng (hyperplane) có (n - 1) chiều trong không gian n chiều của dữ liệu sao cho siêu phẳng này phân loại các lớp một cách tối ưu nhất. Nói cách khác, cho một tập dữ liệu có nhãn (học có giám sát), thuật toán sẽ dựa trên dữ liệu học để xây dựng một siêu phẳng tối ưu được sử dụng để phân loại dữ liệu mới. Ở không

gian 2 chiều thì siêu phẳng này là 1 đường thẳng phân cách chia mặt phẳng không gian thành 2 phần tương ứng 2 lớp với mỗi lớp nằm ở 1 phía của đường thẳng.

- Random Forests (RF):

  RF là thuật toán học có giám sát có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Random Forests giúp hạn chế việc overfitting so với Decision Tree.

- K-Nearest Neighbors (KNN):

  KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu. Hình dưới đây là một ví dụ về KNN trong classification với K = 1. Khi training, thuật toán này không học một điều gì từ dữ liệu training, mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. KNN có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression.

- Logistic Regression (LR):

  LR là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hóa sử dụng hàm tuyến tính (bậc 1). Các tham số của mô hình (hay hàm số) được ước lượng từ dữ liệu.

CHƯƠNG 2: CÁC NGHIÊN CỨU LIÊN QUAN

1. Adapting Transformers for Multi-Label Text Classification (Haytame Fallah, Patrice Bellot, Emmanuel Bruno, Elisabeth Murisasco)

Các mô hình ngôn ngữ được tiền huấn luyện đã chứng tỏ hiệu quả trong phân loại văn bản đa lớp. Nghiên cứu này cải thiện cách tiếp cận này cho phân loại văn bản đa nhãn, một nhiệm vụ đáng ngạc nhiên là ít được khám phá trong vài năm qua mặc dù có nhiều ứng dụng thực tế. Ở đây, tính độc đáo của Haytame Fallah và các nhà nghiên cứu là đề xuất kiến trúc cho các lớp phân loại được sử dụng trên các biến đổi để cải thiện hiệu suất của chúng cho phân loại đa nhãn. Họ đánh giá các phương pháp ngưỡng trên một số biến đổi, bằng cách tính toán ngưỡng riêng cho từng nhãn (IT) hoặc ngưỡng toàn cầu (GCT). Cách thứ nhất bao gồm thêm một tham số để học số lượng nhãn hiện tại cho một ví dụ nhất định (NHA). Cách tiếp cận thứ hai bao gồm thêm một lớp vào các lớp phân loại để học các đặc điểm nhằm lựa chọn các nhãn phù hợp trong khi tránh sử dụng ngưỡng (TL). Việc này được thực hiện trên hai tập dữ liệu tiếng Anh gồm các bài báo trên báo (Reuters-21578 dataset) và các bài báo khoa học (AAPT dataset), sau đó trên một tập dữ liệu đa nhãn mới gồm các tóm tắt bài báo khoa học tiếng Pháp có sẵn công khai. Thống kê kết quả cho thấy hiệu suất đề xuất đã vượt trội so với các phương pháp phân loại văn bản đa nhãn tiên tiến nhất cho các tập dữ liệu được đánh giá và có thể áp dụng cho bất kỳ vấn đề phân loại đa nhãn nào với Precision của mô hình Random Forest đạt 97,18% với dataset Reuter và 94,2% với dataset AAPT.

2. Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning (Emre Deniz , Hasan Erbay and Mustafa Cosar)

Trong phạm vi nghiên cứu này, tác giả tạo một tập dữ liệu đánh giá của khách hàng mới bao gồm các đánh giá của người tiêu dùng Thổ Nhĩ Kỳ để thực hiện phân tích được đề xuất. Tập dữ liệu đã tạo chứa hơn 50.000 đánh giá thuộc ba danh mục khác nhau và mỗi đánh giá có nhiều nhãn tùy theo nhận xét của khách hàng. Sau khi dữ liệu văn bản thô đã được làm sạch và chuẩn hóa, họ chuyển đổi thành một tập đặc điểm của dữ liệu chuỗi số để có thể sử dụng nó trong việc phát triển mô hình phân loại dựa trên văn bản bằng các kỹ thuật trích xuất đặc trưng dựa trên văn bản là TF-IDF,

Word2Vec, GloVe và BERT. Sau đó, họ áp dụng các phương pháp học máy được sử dụng để phân loại nhiều nhãn cho tập dữ liệu. Kết quả cuối cùng cho thấy các mô hình học máy cho các kết quả rất khả thi với Micro Precision 0,9157, Micro Recall 0,8837, Micro F1 Score 0,8925 và Hamming Loss 0,0278 là những phương pháp thành công nhất.

CHƯƠNG 3: XÂY DỰNG BỘ DỮ LIỆU

1. Thu thập dữ liệu

Chúng tôi thu thập dữ liệu từ API của sàn thương mại điện tử Tiki và xây dựng bộ dữ liệu từ 3492 đánh giá của khách hàng đã mua và sử dụng các sản phẩm chăm sóc da và trang điểm. Trong đó:

- Sản phẩm chăm sóc da: 2532 đánh giá.
- Sản phẩm trang điểm: 960 đánh giá.

2. Quy tắc gán nhãn

Để phân tích đánh giá của khách hàng về chất lượng của 1 sản phẩm và thông tin liên quan, chúng tôi chia thành 8 nhãn để phân loại. Trong đó, có 4 nhãn được chọn làm tiêu chí đánh giá chất lượng của 1 sản phẩm:

- Mùi hương: phù hợp với tình trạng của mỗi cá nhân. Ví dụ: da dị ứng với hương liệu.
- Kết cấu: phù hợp với tình trạng của mỗi cá nhân. Ví dụ: da dầu: sử dụng sản phẩm dạng gel hoặc serum.
- Độ bền: phù hợp với nhu cầu sử dụng của mỗi cá nhân. Ví dụ: son lỳ, khả năng kiềm dầu, chống nước.
- Màu sắc: phù hợp với nhu cầu sử dụng của mỗi cá nhân. Ví dụ: lựa chọn son phù hợp với sở thích.

Ngoài ra, chúng tôi dùng 3 nhãn để cung cấp thêm thông tin về sản phẩm:

- Vận chuyển: dịch vụ giao hàng, thời gian giao hàng, người giao hàng.
- Giá
- Đóng gói: bao bì sản phẩm, hình thức đóng gói.

Riêng nhãn dịch vụ vừa để đánh giá chất lượng vừa để cung cấp thông tin:

- Dịch vụ để đánh giá chất lượng: vấn đề thật / giả của sản phẩm.
- Dịch vụ để cung cấp thông tin: về chăm sóc khách hàng, thái độ phục vụ của bên bán sản phẩm.

2.1. Vận chuyển

- Các đánh giá đề cập đến việc giao hàng: hộp bị móp, méo mó; hàng bị ướt.
  Ví dụ: “Tiki Giao hàng móp méo. Ko biết ảnh hưởng tới chất lượng ko. Nhưng vẫn cho 5 sao cho sản phẩm tem mác đầy đủ. Hàng chính hãng”.

- Đề cập đến thời gian giao hàng

  Ví dụ: “Tiki giao hàng nhanh thần tốc chiều nay đặt 10h ngày mai có, quá ưng dịch vụ giao hàng”.

  “Đóng gói siêu siêu chắc chắn. Giao nhanh kiểm dầu tốt, da dầu dưỡng ẩm không kỹ còn có cảm giác hơi khô”.
- Đề cập đến người giao hàng

  Ví dụ: “Gói kỹ, giao nhanh, màu đẹp tự nhiên (theo quan điểm của mình), lâu trôi (mình dùng để tô mày không à). Shipper nhiệt tình thân thiện, Thanks all”.

2.2. Giá

- Các đánh giá đề cập đến mức giá của sản phẩm

  Ví dụ: “Dùng cũng ổn trong tầm giá, sau khi tẩy tế bào chết rồi bôi lên môi trơn mềm.đáng giá mua”.

  “Son rẻ mà không thua kém gì mấy son dưỡng ẩm đắt tiền, mình mua hẳn 4 cây về dùng dần”.
- Đề cập đến giá cụ thể của sản phẩm

  Ví dụ: “Giá sale rẻ chỉ 9k 1 cây, thành phần lành tính nên khá yên tâm, tuy chưa có hiệu quả và tác dụng j nổi bật”.

2.3. Đóng gói

- Các đánh giá liên quan đến hình thức đóng gói của bên bán sản phẩm

  Ví dụ: “shop gói hàng cẩn thận giao hàng nhanh, màu son rất đẹp, rất ưng ý”.
- Đề cập đến bao bì sản phẩm

  Ví dụ: “Hỏng nút ấn, bao bì khi nhận được hàng đã bị xé ra!”.

  “Bao bì đẹp mắt, chắc chắn”.

2.4. Dịch vụ

- Các đánh giá đề cập đến dịch vụ chăm sóc khách hàng, thái độ phục vụ của bên bán sản phẩm

  Ví dụ: “Mình đặt mua 2 serum B5, shop giao sai hàng thành 2 kem chống nắng, liên hệ để đổi hàng mà chưa thấy phần hồi ”.

  “mình nhận hàng thấy hàng rất ok. chưa dùng nhưng thấy rất thích. shop bán hàng rất lịch sự. cám ơn shop.”.
- Đề cập đến việc gửi hàng của bên bán sản phẩm

Ví dụ: “Sản phẩm giống mô tả nhưng sao mình tìm mãi mà không thấy tem phụ với tem niêm phong nằm ở đâu hết”.

“Giao sai sp cho mình rồi, vui lòng đổi hàng giúp mình”.

- Đề cập đến vấn đề thật giả, tem, nhãn mác của sản phẩm

  Ví dụ: “Hàng fake ko giống như mô tả, ảnh”.

  “Sản phẩm chính hãng, sịn sò, có tem phụ công ty”.

  “Hàng bể nắp, rách tem, ko đc bọc, yêu cầu đổi hàng”.

- Đề cập đến hạn sử dụng của sản phẩm

  Ví dụ: “Sài ok lắm ạ, hsd không được lâu tại chai cũng 1 lít”.

  “Hàng chính hãng,giao nhanh 2 giờ,date xa,chất lượng sp tốt.”

- Đề cập đến sự kiện tặng quà của bên bán sản phẩm

  Ví dụ: “Shop ơi, mình mua hàng thấy có quà tặng kèm nhưng khi nhận hàng chỉ có mỗi lọ kem dưỡng thôi”.

  “Sản phẩm thơm, bao bì mới rất đẹp chưa dùng nên chưa biết hiệu quả sao, được tặng thêm 3 tuýp nhỏ rất xinh”.

2.5. Mùi hương

- Các đánh giá đề cập hương liệu của sản phẩm: thơm, mùi hắc, không mùi.

  Ví dụ: “son cute, mùi thơm, giá khá ổn”.

- Đề cập đến mùi hương cụ thể của sản phẩm

  Ví dụ: “Son có mùi mật ong, mềm, dưỡng ẩm tốt. nhưng thời gian giữ ẩm không lâu lắm, tầm 2-3h phải dưỡng lại”.

2.6. Kết cấu

- Các đánh giá đề cập đến độ đậm đặc, trạng thái của sản phẩm: lỏng, đặc, tạo bọt, loãng, chất.

  Ví dụ: “Tuyệt vời, da mình hỗn hợp thiên dầu xài quá đã ko bóng dầu, độ kem đặc nhưng chống nắng tốt ko bị đen khi đi chơi hai ngày liên tục”.

  “Hạn dụng lâu dài nhưng kem loãng lắm không giống chai từng sài.”.

  “Giao hàng đúng hẹn, chất son đẹp, xịn”.

- Đề cập đến trạng thái khi sử dụng trên da: thấm nhanh, thoáng da, mịn da, bết dính, thẩm thấu, vón cục (mủn), dễ tán, mượt, mướt.

Ví dụ: “mùi thơm dễ chịu, chất kem thấm nhanh tuy nhiên hơi ít so với giá tiền. Sẽ mua lại”.

“Cấp ẩm cực tốt, đã mua 2 chai 30ml để sử dụng, da căng bóng mịn thấy rõ, nay mua chai 60ml cho tiết kiệm”.

“Mang tiếng kem sữa. Mà ko trắng cứ ngà ngà. Ko giống hàng chính hãng. Bôi lên mặt ko thấm dính dính từ sáng đến chiều cũng dính”.

“ko hợp với da mình bôi lên thấy có mủn trắng trên mặt”.

“okla nhé. Hộp đẹp, phấn mướt. Mỗi tội m chọn màu hơi tối”.

2.7. Độ bền

- Các đánh giá đề cập đến khả năng kiềm dầu, chống nước của sản phẩm: kiềm dầu, bóng dầu, đổ dầu, đổ mồ hôi, chống nước.

  Ví dụ: “Lớp nền mịn, tệp da, không bóng dầu. Màu đẹp tự nhiên.”.

  “sản phẩm ổn áp lắm hợp với da dầu mụn. Sử dụng xong mặt ko còn tiết dầu nhiều thấy rõ. Nhưng nhớ kèm theo bước dưỡng ẩm ạ, vì có BHA ấy”.
- Đề cập đến khả năng giữ lớp sản phẩm trên da: lỳ, phai, trôi, bám, giữ (ẩm), lem.

  Ví dụ: “Màu son thực tế như ảnh dưới. Màu lì hơi khô môi”.

  “Son đẹp mùi thơm chất son mịn lâu phai”.

  “bút kẻ mắt ra mực đen, lâu trôi mỗi tội đầu hơi to xíu còn mascara thì cũng ổn, làm cong mi đi cả ngày không bị lem”.

2.8. Màu sắc

- Các đánh giá đề cập chung đến màu của sản phẩm

  Ví dụ: “Lớp nền bền. Đều màu, không bị trôi hay mốc khi mồ hôi ra, đánh lên rất mịn da”.

  “Màu đẹp, không khô môi, nhưng nếu in màu ngoài vỏ để dễ tìm màu thì ok hơn, mình mua 3 cây mà 3 vỏ đều y chang nhau, ko phân biệt được màu nào hết”.
- Đề cập đến màu cụ thể của sản phẩm

  Ví dụ: “Sao tôi đặt đỏ cam lại giao màu cherry. Tôi thấy nhiều người phản ánh giờ tôi cũng bị. Nếu hết hàng phải báo khách chứ. Bán hàng tào lao vậy”.

Ví dụ: “Lớp nền bền. Đều màu, không bị trôi hay mốc khi mồ hôi ra, đánh lên rất mịn da”.

“Màu đẹp, không khô môi, nhưng nếu in màu ngoài vỏ để dễ tìm màu thì ok hơn, mua 3 cây mà 3 vỏ đều y chang nhau, ko phân biệt được màu nào”.

- Đề cập đến màu cụ thể của sản phẩm

  Ví dụ: “Sao tôi đặt đỏ cam lại giao màu cherry . Đọc cmt tôi thấy nhiều người phản ánh giờ tôi cũng bị . Nếu hết hàng phải báo khách. Bán hàng tào lao vậy”.

3. Đánh giá độ đồng thuận

Chúng tôi đã chia bộ dữ liệu ra làm 2 bộ và tiến hành gán nhãn (2 người gán 1 bộ). Sau đó, dùng độ đo Cohen Kappa score để so sánh kết quả gán nhãn của 2 người.

$$kappa(\kappa) = \frac{P_o - P_e}{1 - P_e}$$

Trong đó:

- Po là tỷ lệ đồng ý thực tế giữa các người đánh giá.
- Pe là tỷ lệ đồng ý ngẫu nhiên.

Kappa (k) được tính bằng cách lấy sự khác biệt giữa tỷ lệ đồng ý thực tế và tỷ lệ đồng ý ngẫu nhiên và chia cho 1 - tỷ lệ đồng ý ngẫu nhiên. Tỷ lệ đồng ý ngẫu nhiên là tỷ lệ đồng ý mà ta có thể mong đợi nếu các người đánh giá chỉ đang đoán mò. Nó được tính bằng cách lấy tổng số khả năng có thể có cho các người đánh giá và chia cho số người đánh giá.

Kết quả đánh giá độ đồng thuận của 2 bộ lần lượt các nhãn là *vận chuyển, giá, đóng gói, dịch vụ, mùi hương, kết cấu, độ bền, màu sắc* như sau.

Bộ 1:

```
0.9161539574999007
0.8595547268976959
0.9045798595964992
0.5672173984049647
0.959594473497666
0.5466047543898744
0.7901811345915211
0.9252680083387869
```

Hình 3.3.1. Kết quả đánh giá độ đồng thuận của bộ 1.

Bộ 2:

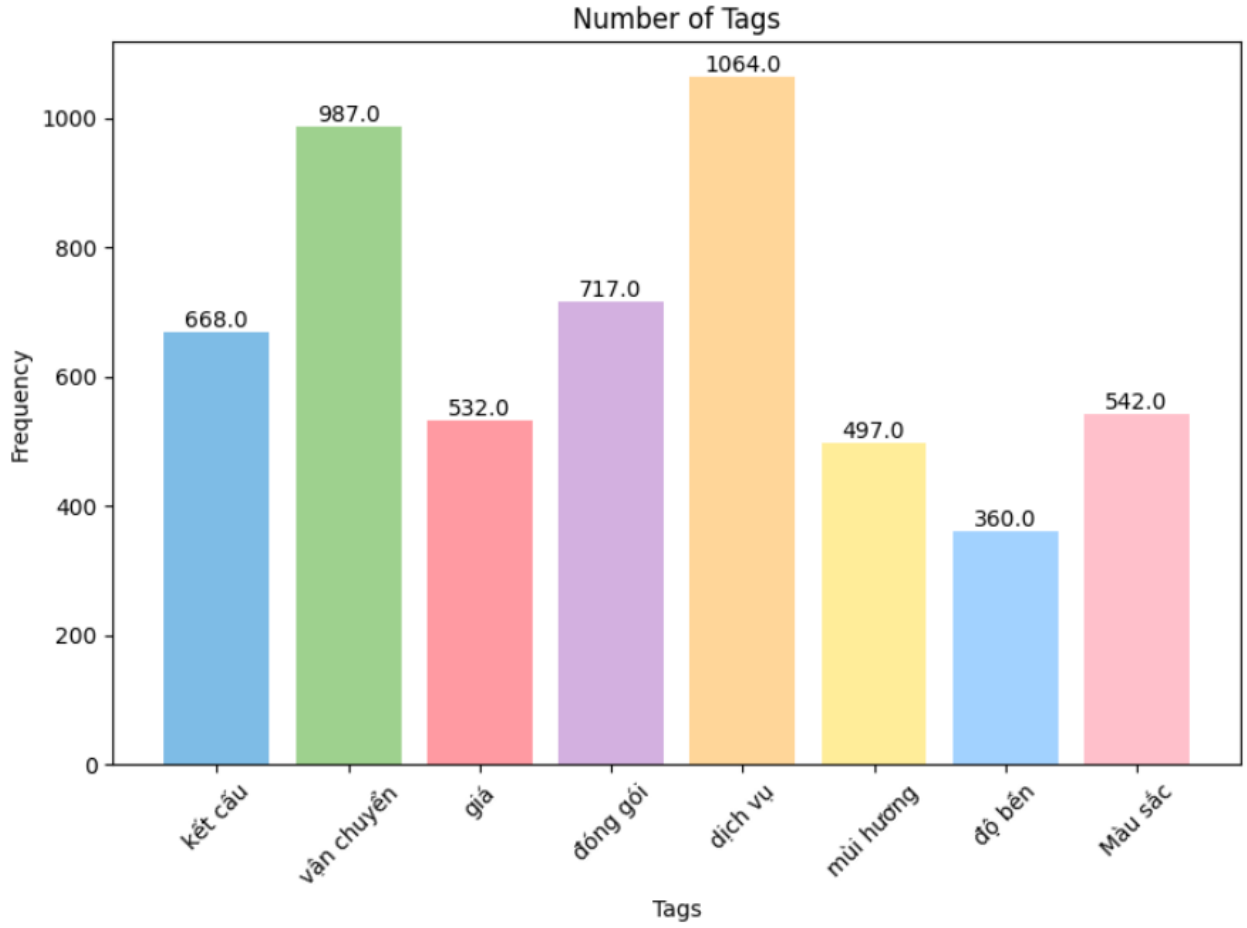
```
0.9649530267090077
0.9159145601171536
0.9376093530345784
0.8253425340656684
0.9819887259833181
0.813787894653587
0.3062173155142358
0.8765155131264917
```

Hình 3.3.2. Kết quả đánh giá độ đồng thuận của bộ 2.

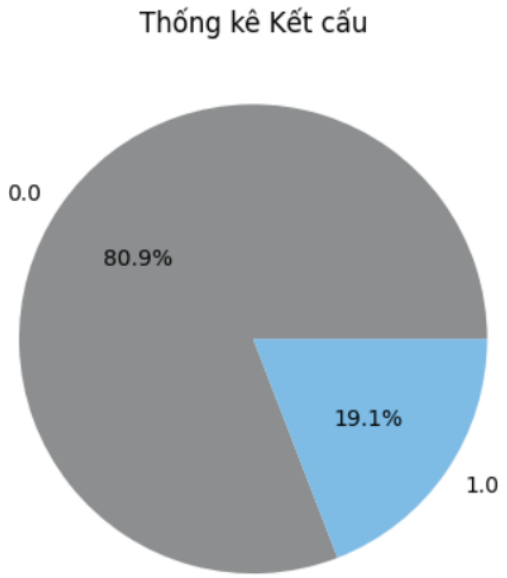
Giá trị độ đo Cohen Kappa càng cao, càng cho thấy sự đồng thuận giữa các người đánh nhãn. Kết quả cho thấy, đa phần các nhãn đều đạt độ đồng thuận cao. Tuy nhiên, do số lượng nhãn khác nhau khá lớn và dữ liệu bị nhập nhằng, nên vẫn có một vài nhãn có sự đồng bộ chưa cao. Để cải thiện tình trạng này, đối với các nhãn có điểm từ 0.3 - 0.5, chúng tôi tiến hành gán nhãn lại. Đối với các nhãn từ 0.8 chúng tôi tìm ra các nhãn khác nhau và thống nhất lại giữa những người gán nhãn về cách đánh giá các nhãn. Cuối cùng thu được bộ dataset hoàn chỉnh.

4. Trực quan hóa dữ liệu

Để có cái nhìn tổng quan và rõ ràng hơn về số lượng các nhãn, chúng tôi tiến hành trực quan hóa dữ liệu và thu được kết quả dưới đây:

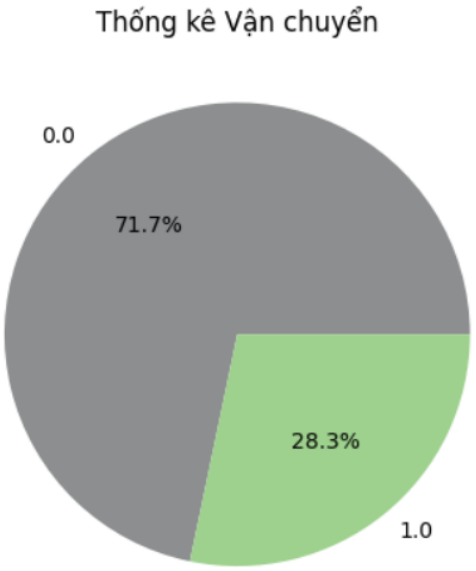


Hình 3.4.1. Kết quả thống kê tổng các nhãn của bài toán.



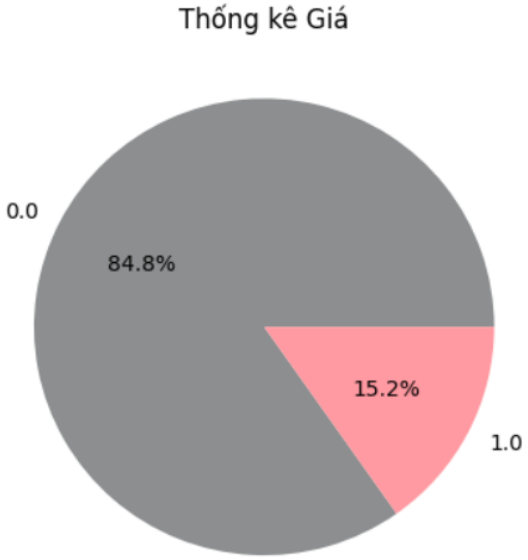

```
0.0    2824
1.0     668
```

Hình 3.4.2. Thống kê nhãn kết cấu.



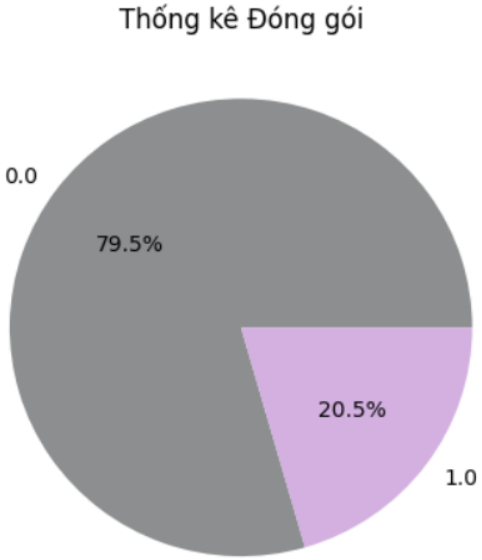

```
0.0    2505
1.0     987
```

Hình 3.4.3. Thống kê nhãn vận chuyển.



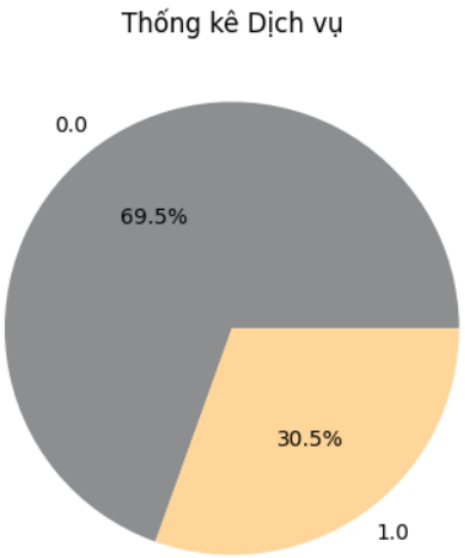

```
0.0    2960
1.0     532
```

Hình 3.4.4. Thống kê nhãn giá.



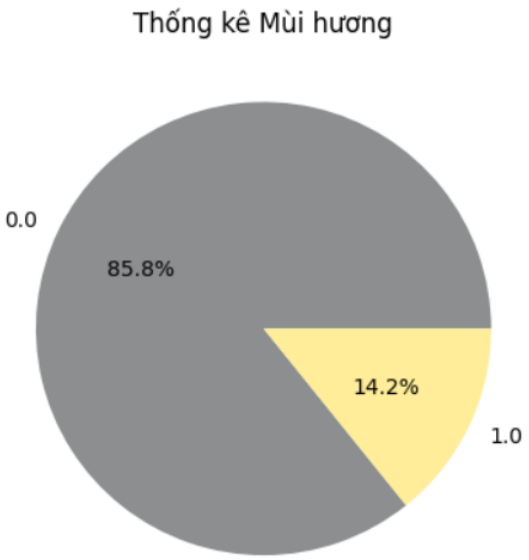

```
0.0    2824
1.0     668
```

Hình 3.4.5. Thống kê nhãn đóng gói.



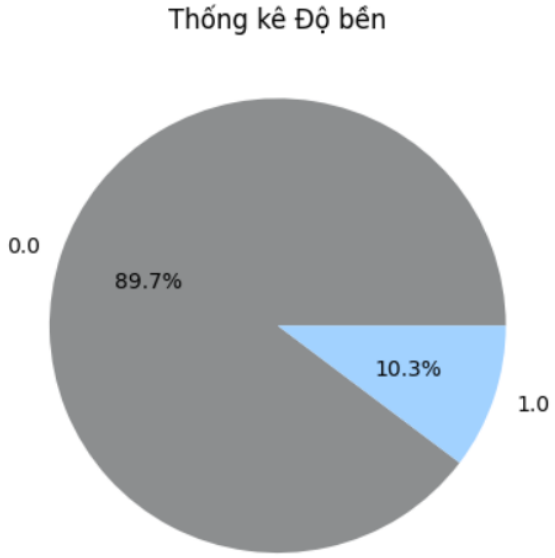

```
0.0    2428
1.0    1064
```

Hình 3.4.6. Thống kê nhãn dịch vụ.



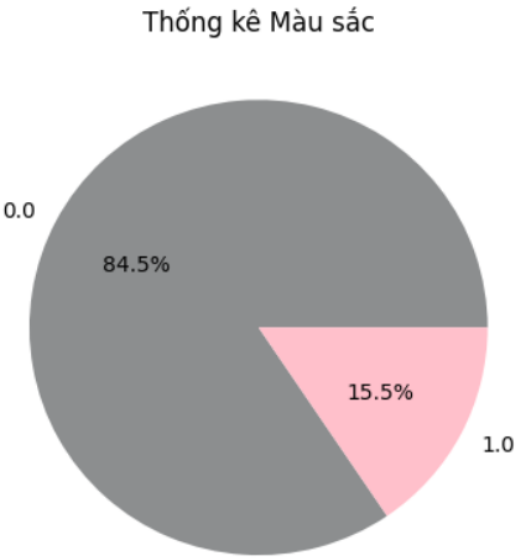

```
0.0    2995
1.0     497
```

Hình 3.4.7. Thống kê nhãn mùi hương.




```
0.0    3132
1.0     360
```

Hình 3.4.8 Thống kê nhãn độ bền.




```
0.0    2950
1.0     542
```

Hình 3.4.9. Thống kê nhãn màu sắc.

CHƯƠNG 4: XỬ LÝ DỮ LIỆU VÀ TRÍCH XUẤT ĐẶC TRƯNG

1. Tiền xử lý dữ liệu

Chúng tôi đã thực hiện các bước như sau để làm sạch dữ liệu, đồng thời tiến hành thử nghiệm sự ảnh hưởng của việc tiền xử lý đến với kết quả cuối cùng (Chương 5: Mục 5).

- Xóa ký tự đặc biệt, ký tự cảm xúc (icon) và xóa các ký tự không cần thiết trong dataset.
  Ví dụ: "[,./)*^]"
- Đưa toàn bộ dữ liệu về chữ viết thường.
- Xóa các ký tự cố ý viết dài.
  Ví dụ: "sản phẩm okeeeeeeeeeeee lắmmmmm nheee" -> "sản phẩm ok lắm nhe".
- Chuẩn hóa unicode.
- Chuẩn hóa dấu câu cho đúng vị trí ví dụ úy òa thay vì uý oà.
- Chuẩn hóa teencode (1 vài trường hợp các từ viết tắt).
  Đối với teencode chúng tôi thu thập từ bộ từ điển có sẵn trên github [] , đồng thời, thu thập thêm các từ viết tắt thường xuyên xuất hiện trên bộ dataset.
  Ví dụ: hsd: hạn sử dụng; kcn: kem chống nắng; srm: sữa rửa mặt; ckđ: che khuyết điểm, tbc: tế bào chết, ...
- Tách từ
- Xóa khoảng trắng thừa.
- Lọc stop word dựa trên số lượng từ xuất hiện nhiều nhất trong bộ dataset.

2. Trích xuất đặc trưng

2.1. TF-IDF có smooth

Phương pháp TF-IDF (Term frequency-inverse document frequency)[]. TF-IDF tính toán độ quan trọng cho mỗi từ trong văn bản nằm trong một tập văn bản lớn.

- Tf- term frequency: Tính toán số lần xuất hiện của một từ trong văn bản.
- IDF - Inverse Document Frequency: Ước lượng độ quan trọng của từ đó trong văn bản. Ví dụ như các từ "và, có, là, của, à,.." xuất hiện nhiều trong văn bản

nhưng không đem lại nhiều ý nghĩa, lúc này trọng số của các từ này sẽ thấp.

$$log\frac{1+n_d}{1+\mathrm{df}(d,t)}+1$$

TF-IDF được tính như sau:

TF-IDF = TF * IDF

Như vậy, giá trị của TF-IDF sẽ tỷ lệ thuận đối với độ quan trọng của từ đó trong toàn bộ tập văn bản. Đối với các từ có điểm cao, đồng nghĩa với mức độ đặc biệt và hiếm gặp trên bộ văn bản. Trái lại, nếu có giá trị thấp thì các từ đó có thể là các từ xuất hiện đại trà trong toàn bộ tập dữ liệu.

2.2. Word2Vec

Word2Vec là một phương pháp biểu diễn từ dựa trên mạng neural nhân tạo, mục tiêu là nắm bắt ý nghĩa của các từ; nó được đề xuất bởi Thomas Mikolov vào năm 2013 [4]. Word2Vec có 2 phương pháp là skip-grams và CBOW. Trong bài toán này, chúng tôi chỉ sử dụng phương pháp skip-grams. Đầu tiền ta sẽ lựa ra ngẫu nhiên các từ làm bối cảnh (context). Dựa trên từ bối cảnh, các từ mục tiêu (target) sẽ được xác định nằm trong phạm vi xung quanh từ bối cảnh.

2.3. Bag of Word

Bag of Words (BOW) là một phương pháp để trích xuất các đặc điểm từ các dữ liệu văn bản. Các đặc điểm này có thể được sử dụng để đào tạo các thuật toán học máy. Nó tạo ra một kho từ vựng chứa tất cả các từ duy nhất có trong tất cả các dữ liệu văn bản trong tập huấn luyện. Hay nói cách khác, đó là một tập hợp bao gồm các cặp giá trị key và value, giá trị key là từ duy nhất có trong tập dữ liệu, giá trị value là số lần xuất hiện của từ đó trong câu, và BOW hầu như không quan tâm đến thứ tự xuất hiện của các từ đó.[5]

CHƯƠNG 5: HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

1. Phương pháp huấn luyện

Về cơ bản, có 3 phương pháp để giải quyết bài toán phân loại đa nhãn là:

- Problem Transformation
- Adapted Algorithm
- Ensemble approaches

Ở phạm vi báo cáo lần này, chúng tôi trình bày là sử dụng phương pháp Problem Transformation.

Ở phương pháp này, chúng tôi sẽ thực hiện bằng 3 cách khác nhau:

- Binary Relevance
- Classifier Chains
- Label Powerset

1.1. Binary Relevance

Ở cách này, mỗi nhãn sẽ được xem là một bài toán phân loại riêng biệt. Ví dụ trong trường hợp dưới đây:

X	Y_1	Y_2	Y_3	Y_4
$x^{(1)}$	0	1	1	0
$x^{(2)}$	1	0	0	0
$x^{(3)}$	0	1	0	0
$x^{(4)}$	1	0	0	1
$x^{(5)}$	0	0	0	1

Với X là feature và Y là các labels. bài toán trên sẽ được chia làm 4 bài toán nhỏ riêng biệt (4 labels):

X	Y_1
$x^{(1)}$	0
$x^{(2)}$	1
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	0

X	Y_2
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	1
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_3
$x^{(1)}$	1
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	0
$x^{(5)}$	0

X	Y_4
$x^{(1)}$	0
$x^{(2)}$	0
$x^{(3)}$	0
$x^{(4)}$	1
$x^{(5)}$	1

Sau đó, dùng các thuật toán phân loại để huấn luyện mô hình. Chúng tôi sử dụng thuật toán Naive Bayes, Random Forest, SVC…

Nhược điểm của phương pháp phân loại này là nó không xem xét mức độ tương quan giữa các nhãn.

1.2. Classifier Chains

Ban đầu mô hình được train trên input đầu vào và một nhãn. sau khi huấn luyện xong, nhãn đã được huấn luyện sẽ trở thành input và tiếp tục huấn luyện để dự đoán ra nhãn tiếp theo.

X	y1
x1	0
x2	1
x3	0

Classifier 1

X	y1	y2
x1	0	1
x2	1	0
x3	0	1

Classifier 2

X	y1	y2	y3
x1	0	1	1
x2	1	0	0
x3	0	1	0

Classifier 3

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0

Classifier 4

Đối với cách làm này chúng ta sẽ hạn chế được nhược điểm trước đó vì nó có xét đến độ tương quan giữa các nhãn.

Nhược điểm của phương pháp này là nếu có nhãn thiếu sự tương quan với nhau thì kết quả sẽ không tốt.

1.3. Label Powerset

Cách làm này sẽ xem xét các trường hợp có các nhãn là giống nhau mà gom lại thành một lớp.

X	y1	y2	y3	y4
x1	0	1	1	0
x2	1	0	0	0
x3	0	1	0	0
x4	0	1	1	0
x5	1	1	1	1
x6	0	1	0	0

X	y1
x1	1
x2	2
x3	3
x4	1
x5	4
x6	3

Như ở ví dụ dưới đây, x1 và x4 có nhãn giống nhau nên sẽ được gom thành một lớp, tương tự với x3 và x6.

Nhược điểm của cách làm này là khi bộ dữ liệu lớn và số lượng nhãn nhiều, thì sẽ sinh ra nhiều class, Làm cho độ phức tạp tăng thì điểm sẽ không cao.

2. Các chỉ số đánh giá (Evaluation Metrics)

Phân loại đa nhãn đòi hỏi các kỹ thuật đánh giá khác so với với các kỹ thuật đánh giá phân loại đơn nhãn truyền thống. Trong bài báo cáo này, chúng tôi sử dụng các độ đo Hamming Loss (HL), Micro Averaged Precision (MicroP), Macro Averaged Precision (MacroP), Micro Averaged Recall (MicroR), Macro Averaged Recall (MacroR), Micro F1 Score (Micro F1) và Macro F1 Score (Macro F1).

$$HL = \frac{1}{|N| \cdot |L|} \sum_{l=1}^{L} \sum_{i=1}^{N} Y_{i,l} \oplus X_{i,l}.$$

Hamming Loss (HL) là tỉ lệ nhãn sai trên tổng số nhãn.

$$MicroP = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FPs(c_i)}.$$

Micro-average precision đo lường độ chính xác của mô hình trên từng lớp (class) và sau đó tính trung bình của các độ chính xác này theo tỉ lệ trọng số của từng lớp.

$$Micro\ R = \frac{\sum_{c_i \in C} TPs(c_i)}{\sum_{c_i \in C} TPs(c_i) + FNs(c_i)}.$$

Micro-average recall đo lường độ chính xác của mô hình trên từng lớp (class) chia cho tổng số lần mô hình đã dự đoán đúng cộng với số lần mô hình đã dự đoán sai trên từng lớp.

$$MacroP = \frac{\sum_{c_i \in C} P(D, c_i)}{|C|}.$$

Macro-average precision là trung bình cộng của các precision theo class. Để tìm Macro-average precision, trước tiên chúng ta cần tính precision của từng lớp. Sau đó, giá trị MacroP được tính bằng cách lấy trung bình của tất cả các precision này.

$$MacroR = \frac{\sum_{c_i \in C} R(D, c_i)}{|C|}.$$

Macro-averaged recall là trung bình cộng của cá recall theo class. Để tính Macro, trước tiên cần tính recall của từng lớp. Sau đó, Macro được tính bằng cách lấy trung bình tất cả các recall này.

$$MicroF1 = 2 \cdot \frac{MicroP \cdot MicroR}{MicroP + MicroR}.$$

Micro-averaged F1 score cho biết điểm F1 được tổng hợp từ tất cả các lớp. Micro averaged F1 score được tính bằng tổng của tất cả true positives, false positives và false negatives trên tất cả các nhãn. Sau đó, tính MicroP và MicroR từ tổng này. Cuối cùng, tính giá trị trung bình để thu được Micro-averaged F1 score.

$$MacroF1 = \frac{1}{N} \sum_{i=0}^{N} F1.$$

Macro-averaged F1 score là trung bình của các điểm F1 trên từng nhãn. Macro F1 score được tính bằng cách tính F1 score cho mỗi nhãn và sau đó lấy trung bình của chúng.

3. So sánh kết quả của các phương pháp huấn luyện

3.1. Kết quả

Method	Classifier	Micro F1	Macro F1	Micro P	Macro P	Micro R	Macro R
Binary Relevance	NB	0.67998	0.62674	0.89258	0.91802	0.55636	0.51201
	RF	**0.86726**	**0.84820**	**0.94674**	**0.94678**	**0.78859**	**0.77561**
	KNN	0.21398	0.18181	0.96295	0.68376	0.12679	0.11036
	SVC	0.83023	0.81636	1.0	0.95751	0.75051	0.73722
	LR	0.71783	0.66895	0.99358	0.94588	0.57877	0.53924
Classifier Chains	NB	0.68988	0.65303	0.84250	0.85783	0.60230	0.57352
	RF	**0.85848**	**0.84678**	**0.94791**	**0.94893**	**0.78832**	**0.77955**
	KNN	0.21398	0.18181	0.96571	0.69981	0.12679	0.11036
	SVC	0.82885	0.81462	0.98210	0.95903	0.74601	0.73217
	LR	0.70397	0.65539	0.99499	0.95012	0.55969	0.52116

Label Powerset	NB	0.47738	0.35602	0.89964	0.91638	0.32686	0.26086
	RF	**0.68644**	**0.62493**	**0.92877**	**0.93085**	**0.54600**	**0.50519**
	KNN	0.21398	0.18181	0.94683	0.65222	0.12679	0.11036
	SVC	0.64809	0.59287	0.96652	0.96691	0.49823	0.45651
	LR	0.54890	0.45857	0.95809	0.95509	0.38814	0.33230

Bảng 5.3.1. Kết quả đánh giá cụ thể trên từng mô hình.

Từ kết quả trên, với mỗi phương pháp, chọn ra mô hình có F1-score cao nhất, sử dụng bộ parameter tốt nhất của mô hình đó tiến hành tính toán hàm mất mát, chúng tôi thu được kết quả sau:

Method	**Hamming Loss**	**Micro F1**
Binary Relevance	0.05025	0.86726
Classifier Chains	0.05078	0.85848
Label Powerset	0.10300	0.68644

Bảng 5.3.2. Kết quả đánh giá bằng hàm mất mát và Micro-F1.

3.2. Nhận xét kết quả

Hai độ đo mà chúng tôi quan tâm nhất là F1-score và Hamming Loss. Đầu tiên có thể thấy, Binary Relevance và Classifier Chains là hai phương pháp có điểm cao nhất và gần ngang bằng nhau, Binary Relevance có điểm cao hơn một chút. Kết quả này cho thấy bộ dữ liệu của chúng tôi vừa có sự tương quan và vừa không tương quan giữa nhãn. Có thể hiểu rằng một nhãn có sự tương quan với nhãn này nhưng lại không có sự tương quan với nhãn khác. Ví dụ, trong quá trình gắn nhãn dữ liệu, chúng tôi nhận thấy rằng, khi người dùng đánh giá một vấn đề liên quan đến vận chuyển thì khả năng cao sẽ nhắc đến vấn đề liên quan đến nhãn đóng gói, nhưng lại không có cơ sở để xảy ra khả năng nhắc đến nhãn mùi hương. Thứ hai, trong cả 3 phương pháp thì mô hình Random Forest cho kết quả tốt nhất.

4. So sánh kết quả của các phương pháp trích xuất đặc trưng

Chúng tôi sử dụng phương pháp huấn luyện cho kết quả tốt nhất là Binary Relevance và tiến hành so sánh lại kết quả bằng cách thử nghiệm nhiều cách trích xuất đặc trưng khác nhau.

Method	**Hamming Loss**	**Micro F1**
TF-IDF	0.05025	0.86726
Word2Vec	0.11999	0.62324
Bag_of_word	0.05042	0.86847

Bảng 5.4.1. So sánh kết quả của các phương pháp trích xuất đặc trưng.

Kết luận: 2 phương pháp là TF-IDF và Bag_of_word đều cho kết quả khá tốt. TF-IDF và Bag_of_word biểu diễn từ vựng dựa trên tần số xuất hiện của từ, trong khi Word2Vec xây dựng biểu diễn từ thông qua mô hình neural network và không nhất thiết capture được toàn bộ ngữ cảnh, thông tin ý nghĩa của từ. Với một bộ dữ liệu có kích thước khá nhỏ và mối quan hệ ngữ cảnh giữa các từ khá phức tạp, chẳng hạn như cùng một ý nghĩa nhưng người dùng sử dụng rất nhiều cách diễn đạt khác nhau, điều đó có thể đã gây ra nhiều hạn chế cho phương pháp Word2Vec. Tóm lại, với bộ dữ liệu của chúng tôi, chúng tôi đề xuất sử dụng TF-IDF và Bag_of_word để đạt được kết quả tốt nhất.

5. Sự ảnh hưởng các phương pháp làm sạch văn bản đến kết quả huấn luyện

Chúng tôi đã tiến hành thử nghiệm, sử dụng mô hình huấn luyện tốt nhất từ đó xem xét mức độ ảnh hưởng của các phương pháp làm sạch dữ liệu đã nêu ở trên. Chúng tôi đã lần lượt loại bỏ từng phương pháp và so sánh với điểm số ban đầu để rút ra kết luận.

	Hamming Loss	**Micro F1**
Điểm số tốt nhất	0.05025	0.86726

Phương pháp loại bỏ khỏi tiền xử lý	**Hamming Loss**	**Micro F1**
Xóa kí tự, emoji, khoảng trắng thừa	0.08398	0.82868
Chuẩn hóa cách gõ dấu tiếng Việt	0.08593	0.82400
Xóa các kí tự cố ý viết dài	0.08789	0.81781
Chuẩn hóa teencode	0.07813	0.84127
Chuẩn hóa unicode	0.08203	0.83200
Stopword	0.09375	0.80487
Tách từ	0.08594	0.82677

Bảng 5.5.1. Sự ảnh hưởng các phương pháp làm sạch văn bản đến kết quả huấn luyện.

Kết luận: dễ dàng nhìn thấy được, các phương pháp làm sạch dữ liệu đều ảnh hưởng đến kết quả huấn luyện. Cụ thể, khi loại bất cứ phương pháp nào ra khỏi mô hình, điểm số liền sẽ giảm. Trong đó, việc loại bỏ Stopword gây ra ảnh hưởng lớn nhất và Chuẩn hóa teencode không gây ảnh hướng nhất.

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Chúng tôi đã xây dựng được bộ dữ liệu dành cho bài toán phân phân loại đánh giá của khách hàng trên sàn thương mại điện tử gồm hơn 3500 sample với 8 nhãn liên quan và 2 nhãn là 0 và 1. Chúng tôi đã huấn luyện được bộ dữ liệu và dự đoán các đánh giá sản phẩm của khách hàng ra các nhãn về chất lượng hay thông tin sản phẩm như mục tiêu ban đầu. Chúng tôi đã làm đề tài này qua các bước thu thập và xử lý dữ liệu, trích xuất đặc trưng, huấn luyện mô hình, đưa ra đánh giá. Với kết quả nhận được, phương pháp Binary Relevance cho kết quả tốt nhất và phương pháp Classifier Chains có kết quả gần tương đương, đạt kết quả cao nhất trên mô hình Random Forest.

Tuy nhiên, chúng tôi còn gặp nhiều vấn đề trong dữ liệu như việc mất cân bằng giữa nhãn 0 và 1.

2. Hướng phát triển

Tương lai, chúng tôi sẽ thực hiện kết hợp giữa bài toán này và bài toán multi - class, mục đích là để phân loại các tốt xấu trên từng nhãn mục tiêu và đưa ra nhận xét tổng quan cho người dùng tham khảo. Đồng thời, sẽ sử dụng PhoBert, VisoBert kết hợp cùng các thuật toán Deep Learning cho kết quả được tốt hơn.

TÀI LIỆU THAM KHẢO

1. UK Online Shopping and E-Commerce Statistics for 2017. (2024). Retrieved 8 January 2024, from https://www.nasdaq.com/articles/uk-online-shopping-and-e-commerce-statistics-2017-2017-03-14.
2. By Editorial TeamWriter. 68 Useful eCommerce Statistics You Must Know in 2024. (2024). Retrieved 8 January 2024, from https://wpforms.com/ecommerce-statistics/.
3. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. [1301.3781] Efficient Estimation of Word Representations in Vector Space. (2024). Retrieved 8 January 2024, from https://arxiv.org/abs/1301.3781.
4. Xây dựng chương trình Bag of Words. (2024). Retrieved 8 January 2024, from https://tek4.vn/khoa-hoc/lap-trinh-python-can-ban/xay-dung-chuong-trinh-bag-of-words-bai-tap.