ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

LÊ TRẦN HOÀI ÂN – 18520426

TRẦN QUỐC KHÁNH – 18520908

ĐỒ ÁN CUỐI KỲ

MÔN: HỌC MÁY THỐNG KÊ

LỚP: DS102.K21

PHÂN LOẠI BÌNH LUẬN TRÊN MỘT SẢN PHẨM

CLASSIFICATION OF COMMENTS ABOUT A PRODUCT

SINH VIÊN NGÀNH KHOA HỌC DỮ LIỆU

TP. HỒ CHÍ MINH, 2020

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

LÊ TRẦN HOÀI ÂN – 18520426

TRẦN QUỐC KHÁNH – 18520908

ĐỒ ÁN CUỐI KỲ

MÔN: HỌC MÁY THỐNG KÊ

LỚP: DS102.K21

PHÂN LOẠI BÌNH LUẬN TRÊN MỘT SẢN PHẨM

CLASSIFICATION OF COMMENTS ABOUT A PRODUCT

SINH VIÊN NGÀNH KHOA HỌC DỮ LIỆU

GIẢNG VIÊN HƯỚNG DẪN

TS. NGUYỄN TẤN TRẦN MINH KHANG

ThS. VÕ DUY NGUYÊN

TP. HỒ CHÍ MINH, 2020

LỜI CẢM ƠN

Nhóm xin gửi lời cảm ơn chân thành đến thầy Nguyễn Tấn Trần Minh Khang và thầy Võ Duy Nguyên - giảng viên môn "Máy học thống kê" trong Khoa KH&KTTT đã trang bị cho chúng em những kiến thức, kỹ năng cơ bản cần có để hoàn thành đề tài nghiên cứu này

Tuy nhiên trong quá trình nghiên cứu đề tài, do kiến thức chuyên môn còn hạn chế nên nhóm vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá và trình bày về đề tài. Rất mong nhận được sự quan tâm, góp ý, đánh giá của thầy cô giảng viên để các thành viên trong nhóm có thể phát triển hơn về mặt chuyên môn cho tương lai

Xin chân thành cám ơn!

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Ý nghĩa
1	BOW	Bag-of-Words
2	LR	Logistic Regression
3	NLP	Natural Language Processing
4	SVM	Support Vector Machine
5	NBC	Native Bayes Classifer

TÓM TẮT BÁO CÁO

Các nghiên cứu hiện tại về phân tích ngôn ngữ sắc thái bình luận thường được định hướng theo các nhiệm vụ phân loại đơn và đơn ngữ. Trong bài báo cáo này, chúng tôi trình bày một bộ dữ liệu phân tích sắc thái bình luận trên một sản phẩm âm nhạc "Justin Beiber – What do you mean" với 1056 điểm dữ liệu tương ứng với 1056 bình luận của người dùng và sử dụng nó kiểm tra các phương pháp học máy hiện đại nhằm tạo ra một mô hình có thể phân loại được cảm xúc của đánh giá dựa trên bộ dữ liệu đã xây dựng. Chúng tôi đánh giá tập dữ liệu của này trong các cài đặt phân loại khác nhau mà cụ thể ở đây là Logistic Regression, Support Vector Machine và MultinomialNB, sau đó chúng tôi thảo luận cách tận dụng các thông số mô hình và chú thích của chúng tôi để cải thiện sự phát hiện và phân loại ngôn từ sắc thái bình luận.

Hình 1. *Music video "What do you mean" - Justin Beiber*[1]

[1] Music video "What do you mean" – Justin Beiber - https://www.youtube.com/watch?v=DK_0jXPuIr0

CHƯƠNG 1. GIỚI THIỆU

1.1. Giới thiệu tổng quan về đề tài

Với lượng dữ liệu văn bản được mở rộng trên các nền tảng truyền thông xã hội khác nhau mà đặc biệt ở đây là trên nền tảng của Youtube, các bộ lọc hiện tại chưa cung cấp đủ các công cụ để phân loại sắc thái bình luận: nhận biết, ưu tiên các bình luận mang sắc thái tích cực, trung tính và ngăn chặn sự lây lan của các bình luận tiêu cực, ngôn từ kích động thù địch. Hầu hết người dùng Youtube tham gia vào đánh giá, nhận xét các sản phẩm ở đây với các tác động của cảm xúc của cá nhân lẫn sự tác động của những người dùng khác.

Đối với xử lý, phân loại ngôn ngôn ngữ cảm xúc như một nhiệm vụ nhị phân có thể không đủ để kiểm tra động lực và hành vi của người dùng thúc đẩy nó và mọi người sẽ phản ứng với nó như thế nào.

- Đối với các bình luận mang cảm xúc tích cực: Các bình luận mang sắc thái tích cực hướng tới rất nhiều chủ đề và các nhóm đối tượng khác nhau. Tuy nhiên, chủ yếu trong số đó là các chủ đề về: sự hân hoan, sự phấn khích, hạnh phúc, vui sướng, vui lòng, ….
- Đối với các bình mang cảm xúc trung tính: Các bình luận liên quan đến sản phẩm mang cảm xúc trung tính, không chứa đựng các nội dung biểu thị sắc thái cảm xúc tích cực hoặc tiêu cực.
- Đối với các bình luận mang tính tiêu cực: Ví dụ, các bình luận đáng ghét được trình bày bên dưới cho thấy tính tiêu cực hướng vào các mục tiêu khác nhau, có hoặc không sử dụng ngôn ngữ nhếch nhác và tạo ra một số loại phản ứng.

Chúng tôi tin rằng, để cân bằng giữa sự thật và tính chủ quan, có ít nhất ba khía cạnh quan trọng trong phân tích cảm xúc. Do đó, các chú thích của chúng tôi chỉ ra:

- Văn bản là trực tiếp hay gián tiếp;
- Thuộc tính dựa trên đó nó đề cập tới một cá nhân hoặc một nhóm người;
- Các nhà chú thích cảm thấy thế nào về nội dung của nó trong phạm vi từ tình cảm tiêu cực đến trung tính.

Chúng tôi tin rằng lược đồ chú thích đa khía cạnh với các loại cảm xúc của chúng tôi sẽ cung cấp một cái nhìn tương đối sâu sắc và có giá trị về quá trình nhận dạng, phân loại các cảm xúc của các đánh giá, bình luận của người dùng.

Chúng tôi thành lập một nhóm nhà chú thích để gán nhãn cho khoảng 9.000 bình luận được thu thập từ ca khúc "Justin beiber – What do you mean" dựa trên các khía cạnh được đề cập ở trên và coi mỗi khía cạnh là một nhiệm vụ dự đoán. Chúng tôi so sánh các cài đặt mô hình học máy (Machine Learning) Logistic Regression, Support Vector Machine và MultinomialNB tương ứng. Sau đó, chúng tôi báo cáo kết quả hiệu suất của các cài đặt khác nhau và thảo luận về cách mỗi tác vụ ảnh hưởng đến các tác vụ còn lại. Chúng tôi phát hành bộ dữ liệu và mã cho cộng đồng để mở rộng công việc nghiên cứu về phát hiện và phân loại cảm xúc của các bình luận.

1.2. Mục tiêu đề tài:

Mục tiêu đặt ra của bài toán là xây dựng một bộ dữ liệu các bình luận về một sản phẩm âm nhạc trên nền tảng Youtube. Từ đó, áp dụng các mô hình học máy hiện đại để phân loại cảm xúc (emotion classification) các bình luận của người dùng Youtube về bài hát "Justin Beiber – What do you mean" bằng cách phân loại các bình luận theo ba loại cảm xúc: tích cực (positive), tiêu cực (negative), trung tính (neutral).

1.3. Các vấn đề liên quan

Trong quá trình tiến hành xây dựng bộ dữ liệu, xây dựng các mô hình phân loại, bên cạnh việc phát hiện các bình luận mang sắc thái tích cực và trung tính thì việc phát hiện và xử lý các bình luận mang sắc thái tiêu cực lại đi kèm nhiều khó khăn hơn. Có rất ít sự đồng thuận về sự khác biệt giữa lời nói thô tục và lời nói căm thù và, làm thế nào để xác định điều sau (Schmidt và Wiegand, 2017). Những lời xỉ vả không phải là một chỉ số rõ ràng của lời nói căm thù và có thể là một phần của cuộc trò chuyện không gây khó chịu, trong khi một số ý kiến xúc phạm nhất có thể đến dưới dạng ẩn dụ tinh tế hoặc châm biếm (Malmasi và Zampieri, 2018). Do đó, không có từ vựng chú thích nào của con người hiện rõ ràng cho thấy sự hiện diện của ngôn từ kích động thù địch. Điều này cho thấy thô tục không phải là một chỉ số rõ ràng về sự hiện diện của ngôn từ kích động thù địch. có thể ghét lời nói thô thiển và ồn ào (Nobata et al., 2016).

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Chúng tôi báo cáo và thảo luận về kết quả của các nhiệm vụ phân loại: (1) tính trực tiếp của bài phát biểu, (2) loại cảm xúc của bình luận, (3) thuộc tính mục tiêu.

2.1. Giới thiệu mô hình

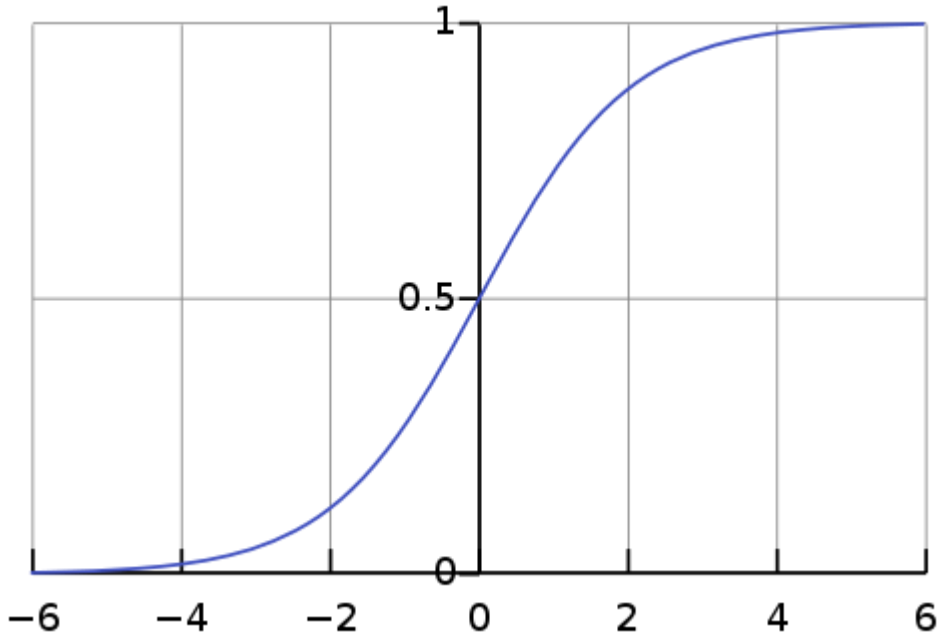
Chúng tôi thực hiện so sánh các thuật toán sử dụng từ bag-of-words (BOW) như các thuộc tính trên mô hình hồi quy Logistic (LR), Support Vector Machine (SVM) và MultinomialNB.

2.2. Cơ sở lý thuyết

2.2.1. Logistic Regression

2.2.1.1. Khái niệm

Hồi quy logistic[2] [1] là phân tích hồi quy thích hợp để tiến hành khi biến phụ thuộc là nhị phân (nhị phân). Giống như tất cả các phân tích hồi quy, hồi quy logistic là một phân tích dự đoán. Hồi quy logistic được sử dụng để mô tả dữ liệu và để giải thích mối quan hệ giữa một biến nhị phân phụ thuộc và một hoặc nhiều biến độc lập danh nghĩa, thứ tự, khoảng hoặc tỷ lệ độc lập.



Hình 2.1. *Nền tảng mô hình hồi quy Logistic*

[2] Machine Learning cho người mới bắt đầu: https://viblo.asia/p/machine-learning-cho-nguoi-moi-bat-dau-part-2-naQZR1WXKvx

2.2.1.2. Các giả định của mô hình hồi quy Logistic

- Biến phụ thuộc phải có tính chất lưỡng phân.
- Không nên có các ngoại lệ trong dữ liệu, có thể được đánh giá bằng cách chuyển đổi các yếu tố dự đoán liên tục thành điểm số được tiêu chuẩn hóa và loại bỏ các giá trị dưới -3,29 hoặc lớn hơn 3,29.
- Không nên có mối tương quan cao (đa cộng đồng) giữa các yếu tố dự đoán.

Điều này có thể được đánh giá bằng một ma trận tương quan giữa các yếu tố dự đoán. Tabachnick và Fidell (2013) đề xuất rằng các hệ số tương quan dài giữa các biến độc lập nhỏ hơn 0,90, giả định được đáp ứng.

Tại trung tâm của phân tích hồi quy logistic là nhiệm vụ ước tính tỷ lệ cược log của một sự kiện. Về mặt toán học, hồi quy logistic ước tính hàm hồi quy tuyến tính đa định nghĩa là:

$$P = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

2.2.1.2. Các khái niệm liên quan

Overfiting[3]: Khi chọn mô hình cho phân tích hồi quy logistic, một xem xét quan trọng khác là sự phù hợp của mô hình. Việc thêm các biến độc lập vào mô hình hồi quy logistic sẽ luôn tăng lượng phương sai được giải thích trong tỷ lệ cược log (thường được biểu thị bằng R^2). Tuy nhiên, việc thêm ngày càng nhiều biến vào mô hình có thể dẫn đến quá mức, điều này làm giảm tính tổng quát của mô hình ngoài dữ liệu mà mô hình phù hợp.

Báo cáo R2[4]: Nhiều giá trị giả R2 đã được phát triển cho hồi quy logistic. Chúng nên được giải thích hết sức thận trọng vì chúng có nhiều vấn đề tính toán khiến chúng cao hoặc thấp một cách giả tạo. Một cách tiếp cận tốt hơn là trình bày bất kỳ sự tốt đẹp của các bài kiểm tra phù hợp có sẵn.

[3] Machine Learning cơ bản – Overfiting: https://machinelearningcoban.com/2017/03/04/overfitting/
[4] R2 (R square) score: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

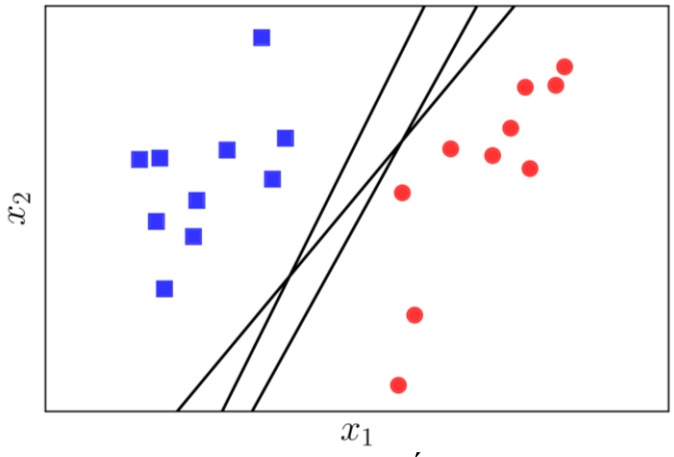
2.2.2. Suport Vector Machine (SVM)

2.2.2.1. Khái niệm

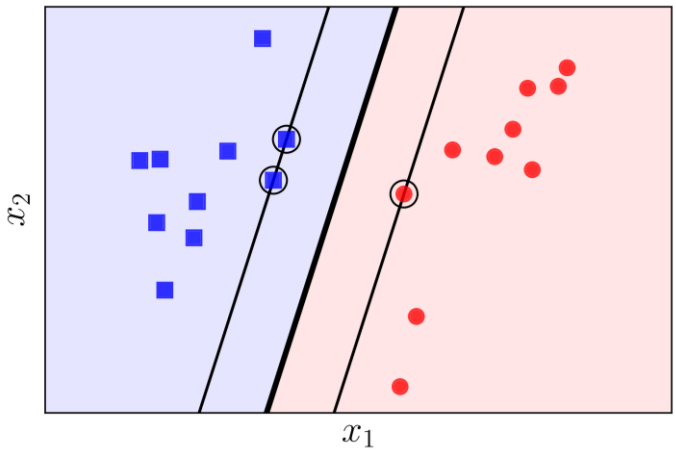
Support Vector Machine[5] [2] là một trong các thuật toán phân loại được sử dụng phổ biến nhất trong Machine Learning.

Đây là một thuật toán phân loại tuyến tính. Tức là sau khi huấn luyện, ta thu được các siêu phẳng phân chia các lớp dữ liệu với nhau, giống như thuật toán Logistic Regression.

Mục tiêu của thuật toán này không những phân chia được các lớp dữ liệu với nhau, mà còn tìm cách để tối đa khoảng cách giữa đường phân chia với các điểm dữ liệu giữa các lớp (maximum margin).



Hình 2.2. *Mô hình phân loại tuyến tính thông thường*



Hình 2.3. *Mô hình phân loại SVM*

[5] Machine Learning cơ bản – Support Vector Machine: https://machinelearningcoban.com/2017/04/09/smv/

Thuật toán SVM sẽ tìm một số vector đặc biệt (gọi là support vectors).

Mô hình (Model) dự đoán (predict) kết quả đầu ra của những điểm dữ liệu mới dựa trên các vector đặc biệt này.

2.2.2.2. Điểm đặc biệt của Suport Vector Machine

- Hầu hết các thuật toán Machine Learning khác đều phân chia dữ liệu dựa trên các điểm dữ liệu đặc trưng nhất của lớp dữ liệu đó.

- Trong khi đó, Support Vector Machine phân chia dữ liệu dựa trên các điểm dữ liệu dễ gây nhầm lẫn nhất giữa các lớp dữ liệu.

2.2.3. MultinomialNB

2.2.3.1. Khái niệm

Mô hình MultinomialNB[6] [3] chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words. Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó.

Ta tính xác suất từ xuất hiện trong văn bản P(xi|y) như sau

$$P(x_i|y) = \frac{N_i}{N_c}$$

Trong đó:

- Ni là tổng số lần từ xi xuất hiện trong văn bản.
- Nc là tổng số lần từ của tất cả các từ x1, …, xn xuất hiện trong văn bản.

Công thức trên có hạn chế là khi từ xi không xuất hiện lần nào trong văn bản, ta sẽ có Ni=0. Điều này làm cho P(xi|y) = 0

4.2.3.2. Cải tiến mô hình

Để khắc phục vấn đề này, người ta sử dụng kỹ thuật gọi là Laplace Smoothing bằng cách cộng thêm vào cả tử và mẫu để giá trị luôn khác 0

[6] Machine Learning cơ bản – Native Bayes: https://machinelearningcoban.com/2017/08/08/nbc/

$$P(x_i|y) = \frac{N_i + \alpha}{N_c + d\alpha}$$

Trong đó:

- α thường là số dương, bằng 1.
- dα được cộng vào mẫu để đảm bảo ∑i=1dP(xi|y) = 1

2.3. Huấn luyện mô hình

Tất cả các phương pháp học máy được so sánh đều sử dụng bộ dữ liệu được phân chia giống nhau với tỉ lệ train: test = 8: 2 và kết quả được báo cáo dựa trên tập test.

Các mô hình sau khi được xây dựng được đánh giá trên tập test sử dụng phương pháp k - fold cross - validation[7] nhằm hạn chế tối thiểu hiện tượng overfiting.

```
# Trainning model with cross validation
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
from sklearn.model_selection import StratifiedShuffleSplit

Y = dataset.iloc[:,1].values
sss = StratifiedShuffleSplit(n_splits=10, test_size=0.2, random_state=0)
sss.get_n_splits(X,Y)
```

***Hình 2.4**. Huấn luyện mô hình với phương pháp K-Fold Cross Validation và test_size = 0.2*

Tại đây, ta sẽ sử dụng các tham số với giá trị mặc định để huấn luyện cho cả 3 mô hình được đưa ra là: Logistic Regression, Support Vector Machine và Multinomial để có thể so sánh và đánh giá khách quan hiệu suất của các mô hình trên**.**

```
#LogisticRegression Model
model_1 = LogisticRegression(random_state=0)
model_1.fit(X_train, Y_train)

#Support Vector Machine Model
model_2 = SVC(kernel='linear')
model_2.fit(X_train, Y_train)

#MultinomialNB Model
model_3 = MultinomialNB()
```

***Hình 2.5.** Huấn luyện cả 3 mô hình với các thông số mặc định*

[7] Cross Validation: https://scikit-learn.org/stable/modules/cross_validation.html

CHƯƠNG 3. BỘ DỮ LIỆU

Trong phần này, chúng tôi trình bày phương pháp thu thập dữ liệu và quy trình chú thích của chúng tôi.

3.1. Thu thập dữ liệu

Việc xem xét sự khác biệt về cảm xúc ở các sản phẩm ở các lĩnh vực khác nhau, những ngôn ngữ khác nhau, những khu vực khác nhau, việc tìm kiếm các thuật ngữ tương đương dẫn đến kết quả khác nhau. Do đó, sau khi tìm kiếm và chọn lọc qua hơn 05 sản phẩm phân biệt, chúng tôi đã tổng kết và quyết định chọn xây dựng và phát triển bộ dữ liệu các bình luận về sản phẩm "Justin Beiber – What do you mean" trên nền tảng Youtube. Trên thực tế, chúng tôi đã bắt đầu thu thập dữ liệu của mình bằng cách sử dụng Web Scraping[8] [4] để có thể lưu trữ các bình luận của bài hát "What do you mean" từ Youtube. Sau đó, chúng tôi tiến hành quan sát và đưa dữ liệu về lưu trữ dưới dạng Comma-Separated Values (.csv) với 01 thuộc tính là "text" – chứa nội dung bình luận của người dùng và tạo thêm 01 thuộc tính mới "Label" – nhãn thể hiện cảm xúc của bình luận *(1: tích cực; 2: trung tính; 3: tiêu cực)* để phục vụ cho các thực nghiệm, phân tích sau này.

cid	text	time	author	votes	photo
Ugzut…	Je laisse mon empreinte... C'E…	1 giờ trước	CuT KuT	0	https://yt3.gg…
Ugy76…	eu amoo essa musica ate hj minha favorita	1 giờ trước	alwaysj	0	https://yt3.gg…
UgwJJ…	Fakin bech jostin	2 giờ trước	Silvia Ortiz	0	https://yt3.gg…
Ugw2C…	This song is LEGANDARY	2 giờ trước	Fynn G	1	https://yt3.gg…
Ugwz5…	Same i cant belive this lol like its very old	3 giờ trước	Brian Gonzalez	0	https://yt3.gg…
Ugy_y…	It was a competition for ZAYN …	3 giờ trước	Albab Hasan	1	https://yt3.gg…
UgzSo…	I haven't heard this song for …	3 giờ trước	Piggy Stuff	1	https://yt3.gg…
UgwBF…	Missed Justin 😔❤	3 giờ trước	Quillen_ xd	1	https://yt3.gg…
Ugz2x…	ALGUÉM ESCUTANDO ESSA MÚSICA EM 2020 EU AMO DMS	3 giờ trước	jose jur…	1	https://yt3.gg…
Ugw-q…	Let's see how many people are …	4 giờ trước	𝖈𝖎𝖘𝖍𝖆	1	https://yt3.gg…
..	Achava que ele era brasileiro	.. .. ..	..	..	.. .. ..

Hình 3.1. *Bộ dữ liệu ban đầu*

[8] Youtube comment downloader: https://github.com/egbertbouman/youtube-comment-downloader

3.2. Thách thức ngôn ngữ

Tất cả các bình luận được chú thích chỉ bao gồm các bình luận gốc, có nội dung được xử lý bằng cách:

3.2.1. Xóa các bình luận spam có thể phát hiện được

Sử dụng công cụ Filter của MS Excel để tiến hành loại bỏ các bình luận có nội dung không liên quan hoặc giữ lại duy nhất 1 bình luận trong số các bình luận giống nhau.

Comment	Label
Aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa	
Aaaaaaaaaaaaa	

Hình 3.2. *Các bình luận spam có thể thấy được*

3.2.2. Xóa các ký tự không phải chữ cái

Chúng tôi tiến hành lọc các ký tự không phải chữ cái vì tin rằng đa số chúng thường không có tác động đến việc thể hiện cảm xúc của người dùng, đa số chúng là do vô tình hoặc cố tình thêm vào để gây độ nhiễu và cần được loại bỏ bằng cách sử dụng hàm re.sub()[9] được cung cấp sẵn trong ngôn ngữ Python.

Comment	Label
==!	

Hình 3.3. *Các ký tự không phải chữ cái*

3.2.3. Xử lý các biểu tượng cảm xúc

Các biểu tượng cảm xúc (emoji) cũng được chuyển đổi thành dạng văn bản để có thể dễ dàng tiến hành các xử lý, thực nghiệm sau này bằng hàm emoji.demojize()[10] trong Python.

Hình 3.4. *Quá trình xử lý các biểu tượng cảm xúc thành văn bản*

[9] Regular expression operations: https://docs.python.org/2/library/re.html
[10] Convert emoji Unicode to text in Python: https://docs.python.org/2/library/re.html

3.2.4. Xử lý văn bản với Bag-of-Words (BOW)

Bag-of-Words[11] [5] là một mô hình dùng để biểu diễn các bình luận thành các vector bằng cách sử dụng Scikit-learns CountVectorizer[12] [6]: Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó.

Ví dụ:

+ Một từ điển có 6 từ gồm: a, am, good, I, not, student.

+ Biểu diễn của các từ là:

Bảng 3.1. *Ví dụ về biểu diễn từ trong mô hình Bag-of-Words*

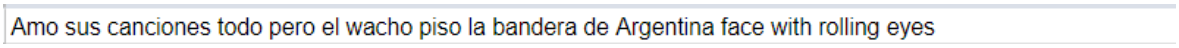
a	[1; 0; 0; 0; 0; 0]
am	[0; 1; 0; 0; 0; 0]
good	[0; 0; 1; 0; 0; 0]
I	[0; 0; 0; 1; 0; 0]
not	[0; 0; 0; 0; 1; 0]
student	[0; 0; 0; 0; 0; 1]

+ Để biểu diễn một câu trong BOW, ta cộng các vector biểu diễn từng từ trong câu lại với nhau:

Bảng 3.2.*Ví dụ về biểu diễn câu trong mô hình Bag-of-Words*

I am a student	[1; 1; 0; 1; 0; 1]
I am a good good student	[0; 1; 2; 1; 0; 1]

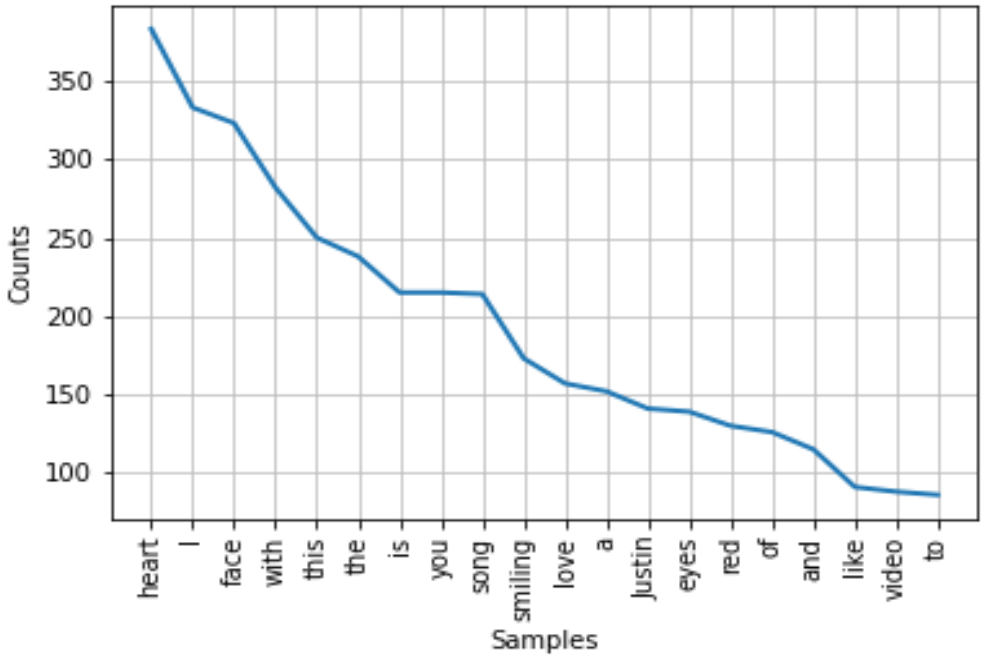
Hơn nữa, chúng tôi nhận thấy việc chuyển đổi mã bằng tiếng Anh trong đó tồn tại một số bình luận cũng chứa các mã thông báo tiếng Hindi, tiếng Tây Ban Nha rất có thể. Do đó, mặc dù chúng tôi đã loại bỏ hầu hết các bình luận này để tránh gây nhầm lẫn cho các chú thích, nhưng những cái còn lại có thể vẫn gây ra độ nhiễu vào dữ liệu.



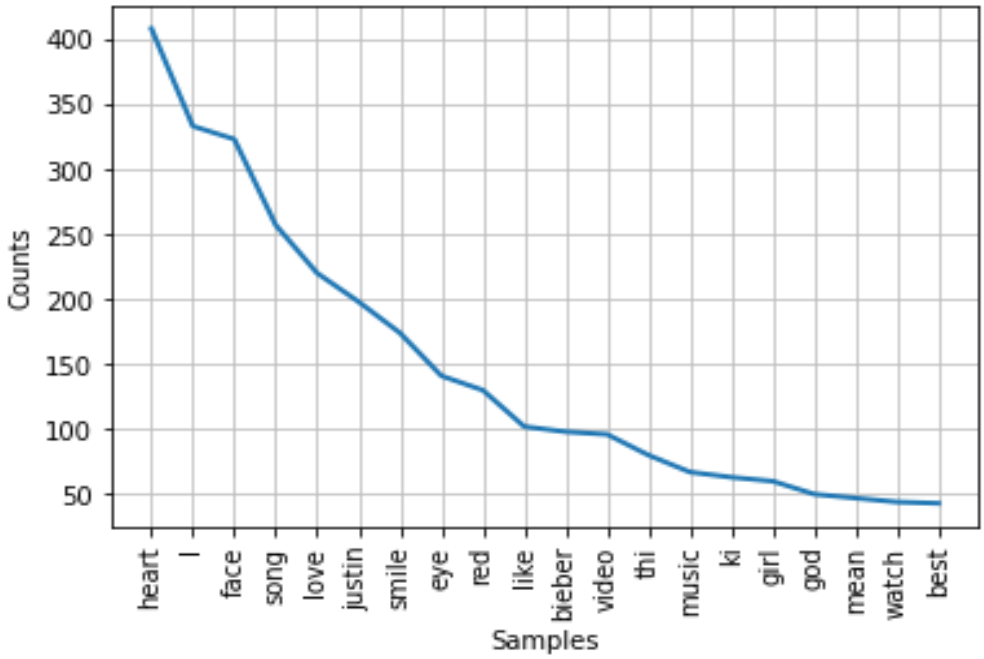
Hình 3.5. *Các bình luận có chứa tiếng Tây Ban Nha, Hindi, ...*

[11] Bag of Words mode in NLP: https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/
[12] About CountVectorizer: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Kết quả là, sau quá trình tiền xử lý, làm sạch dữ liệu, các nhà chú thích đã phải đối mặt với việc thiếu bối cảnh được tạo ra bởi quá trình chuẩn hóa này.



Hình 3.6. *Biểu đồ tần suất xuất hiện của các từ phổ biến (Trước khi xử lý)*



Hình 3.7. *Biểu đồ tần suất xuất hiện của các từ phổ biến (Sau khi xử lý)*

3.3. Quá trình gán nhãn/ chú thích

Chúng tôi dựa vào dư luận chung và kiến thức ngôn ngữ chung để đánh giá cách mọi người xem và phản ứng với sản phẩm này. Chúng tôi cũng đã cung cấp cho các chú thích với một số từ tiếng Anh tiếng lóng mà họ có thể không nhận thức được. Vì sự căng thẳng và tính khó khăn của nhiệm vụ, chúng tôi nhắc nhở các nhà chú thích không nên để ý kiến cá nhân của họ về các chủ đề được thảo luận trong các bình luận ảnh hưởng đến quyết định chú thích của họ.

Hướng dẫn chú thích của chúng tôi đã giải thích thực tế rằng những bình luận này đôi khi biểu hiện nhiều sắc thái cảm xúc xen lẫn nhau. Ở phần này, chúng tôi tập trung vào nhiệm vụ phân tích các bình luận mang sắc thái tiêu cực nhiều hơn vì thực tế rằng những bình luận tiêu cực, xúc phạm và ghét không nhất thiết phải ở dạng thô tục gây ra sự khó khăn nhất định cho công tác gán nhãn. Vì các mức độ phân biệt đối xử khác nhau đối với việc phi nhân cách hóa các cá nhân hoặc nhóm người theo những cách riêng biệt, chúng tôi đã chọn không chú thích các bình luận trong hai hoặc ba lớp. Chẳng hạn, một bình luận phân biệt giới tính có thể thiếu tôn trọng, ghét bỏ hoặc gây khó chịu cho phụ nữ. Nhãn ban đầu của chúng tôi được thành lập phù hợp với các hành vi xã hội phổ biến mà mọi người có xu hướng đối phó. Chúng tôi cũng chọn giải quyết vấn đề về dương tính giả gây ra bởi việc sử dụng sai các từ nhận dạng bằng cách yêu cầu các chú thích ghi nhãn cả các thuộc tính và nhóm mục tiêu.

3.4. Bộ dữ liệu thí điểm:

Chúng tôi ban đầu đặt các mẫu của các bình luận trên hệ thống của Google Docs. Chúng tôi đã cho người xem chú thích bình luận cùng với danh sách các nhãn mô tả

(a) cho dù đó là lời nói trực tiếp hay gián tiếp;

(b) thuộc tính mục tiêu của bình luận;

(c) liệu các chú thích có cảm thấy tức vui vẻ, phấn khích, tức giận,sợ hãi hay không có phản ứng gì về các bình luận.

Mỗi bình luận đã được dán nhãn bởi ba chú thích. Chúng tôi đã cung cấp cho họ văn bản bổ sung các trường để điền vào nhãn hoặc tính từ sẽ

(1) mô tả tốt hơn về bình luận,

(2) mô tả cách họ cảm nhận về nó chính xác hơn.

3.5. Bộ dữ liệu hoàn chỉnh

Bộ dữ liệu cuối cùng là một tập hợp bao gồm 1056 bình luận bằng tiếng Anh với các cảm xúc tích cực, tiêu cực và trung tính cho bài hát "What do you mean". Các nhãn được thiết kế để tạo điều kiện thuận lợi cho việc nghiên cứu mối tương quan giữa người dùng, loại cảm xúc mà nó truyền tải, thuộc tính mục tiêu của nó, nhóm đối tượng mà nó nhắm đến, cách mọi người phản ứng với nó và hiệu suất của việc học đơn nhiệm trên các nhiệm vụ.

Comment	Label
A best song	1
A good song right i m very love it	1
A loveeeeee	1
a quater of the world has seen this video	2
Aaron Carter soon or later is better this is FAKKIN shit	3
Actually he is not good	3
after this nostalgic song sadly was the death of mainstream pop	3
Ahh the memories	2
all of his videos are so cheesy	1
Alove lhis song	1
Always loved you and still do	1

Hình 3.8. *Các quan sát đầu tiên của bộ dữ liệu (10 điểm dữ liệu đầu tiên)*

3.5.1. Code book: Code book mô tả bộ dữ liệu

Bảng 3.3. *Codebook mô tả các thông tin của bộ dữ liệu*

STT	Thông tin	Nội dung
1	Tên bộ dữ liệu	WDYM-EC
2	Nguồn thu thập	https://www.youtube.com/watch?v=DK_0jXPuIr0
4	Kích thước bộ dữ liệu	Bộ dữ liệu gồm 1.056 bình luận đã được gán nhãn chia ra trong 2 bộ: train, test (80%, 20%). Với số lượng câu trong mỗi bộ cụ thể như sau: - Train: có 844 câu - Test: 212 câu

5	Số thuộc tính	Có 02 thuộc tính
6	Thông tin thuộc tính	text: bình luận, kiểu dữ liệu: string label: nhãn cảm xúc, giá trị của mỗi nhãn là một trong 3 giá trị sau: “1”, “2”, “3”.
7	Ý nghĩa các nhãn	Có 3 nhãn cảm xúc là “1,2,3” lần lượt tương ứng với tích cực, trung tính và tiêu cực: - **Tích cực:** là nhãn “1” gán cho những bình luận thể hiện sự hân hoan, sự phấn khích, hạnh phúc, vui sướng, vui lòng, …. VD: “His new music is good I like this song” - **Trung tính:** là nhãn “2” gán cho những bình luận không rõ ràng về nghĩa hoặc không chứa đựng cảm xúc đánh giá về sản phẩm VD: “I do not know what Justin Biber was doing with that girl on the bed” - **Tiêu cực:** là nhãn “3” gán cho những bình luận thể hiện sự không hài lòng, những đòi hỏi, những lời phàn nàn liên quan đến giảng viên, chương trình giảng dạy, cơ sở vật chất… VD: “This is the worst song I have ever heard like you and disgusting get NOOBED nauseated face face vomiting”
8	Tác giả	Lê Trần Hoài Ân, Trần Quốc Khánh

3.5.2. Thống kê dữ liệu

3.5.2.1. Thống kê dữ liệu trên bộ dữ liệu tổng

Bảng 3.4. *Bảng thống kê dữ liệu trên bộ dữ liệu tổng*

STT	Nhãn	Số lượng	Tỉ lệ
1	1	498	47%
2	2	263	25%
3	3	295	28%
Trung bình		352	33.333%
Tổng cộng		1056	100.0%

3.5.2.2. Thống kê dữ liệu trên tập Train

Bảng 3.5. *Bảng thống kê dữ liệu trên tập Train*

STT	Nhãn	Số lượng	Tỉ lệ
1	1	398	47%
2	2	210	25%
3	3	236	28%
Trung bình		281.33	33.333%
Tổng cộng		844	100%

3.5.2.3. Thống kê dữ liệu trên tập Train

Bảng 3.6. *Bảng thống kê dữ liệu trên tập Test*

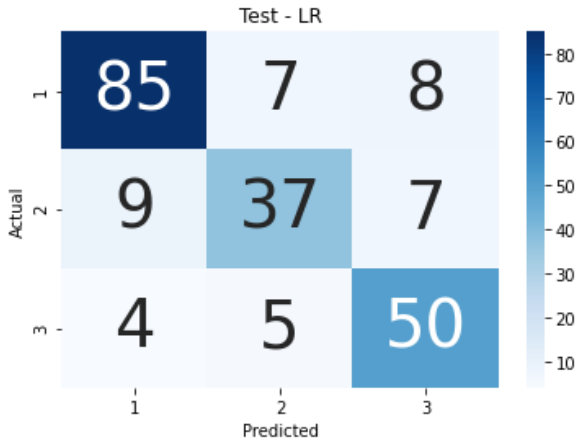
STT	Nhãn	Số lượng	Tỉ lệ
1	1	100	47%
2	2	53	25%
3	3	59	28%
Trung bình		71.33	33.333%
Tổng cộng		212	100%

CHƯƠNG 4. ĐÁNH GIÁ HIỆU SUẤT MÔ HÌNH

Bốn thước đo được chúng tôi sử dụng để đánh giá mô hình là accuracy[13], precision, recall, F1-score đa số đề cập đến ghi nhãn dựa trên nhãn đa số, hồi quy Logistic, Suport Vector Machine, MultinomialNB và đưa ra được kết quả như sau

4.1. Logistic Regression

4.1.1. Confusion matrix và accuracy



Hình 5.1. *Confusion matrix trên tập Test - LR*

Theo ma trận trên, ta thấy:

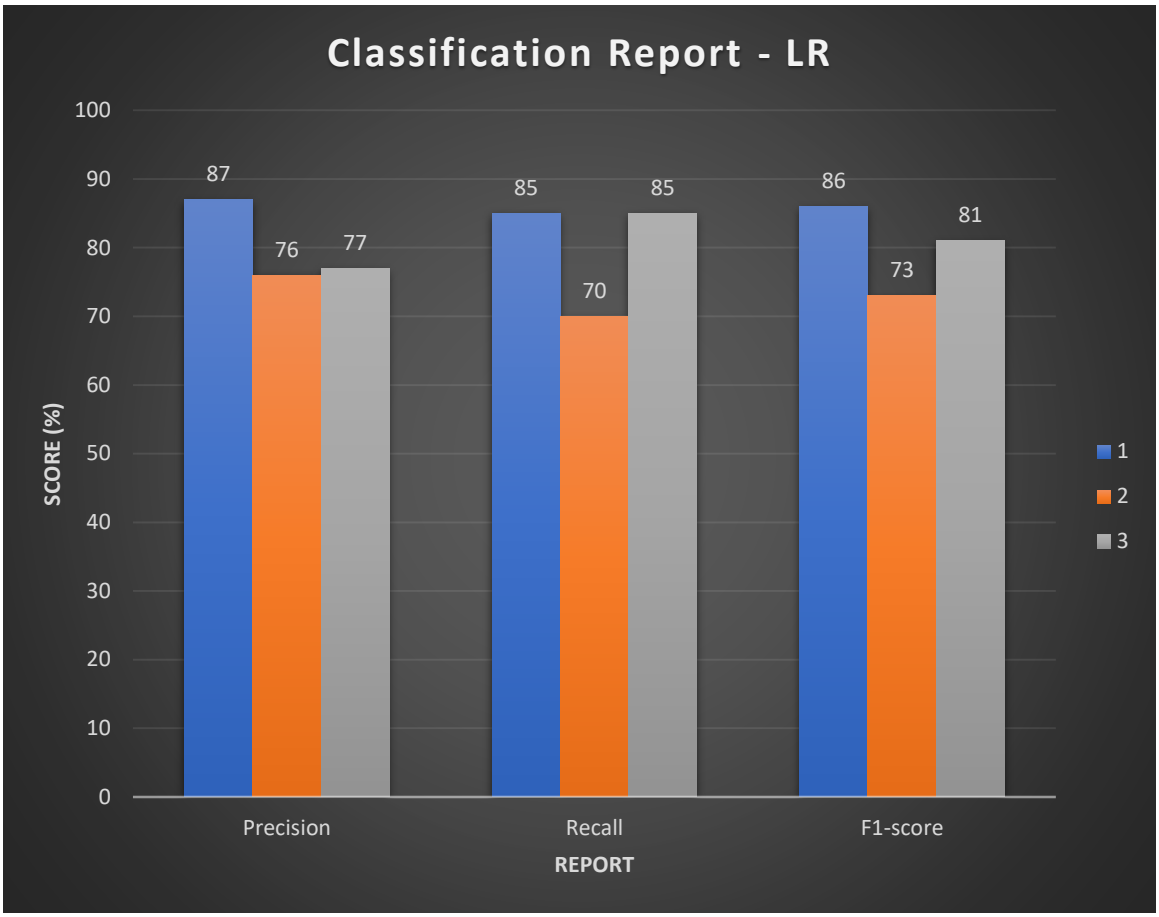
- Số lượng dữ liệu được phân loại đúng là 85 + 37 + 50 = 172 điểm dữ liệu.
- Số lượng dữ liệu phân loại sai là 7 + 8 + 7 + 9 + 4 + 5 = 40 điểm dữ liệu.
- Tỉ lệ điểm dữ liệu phân loại đúng là 172/212 = 0.81132%

4.1.2. Classification report

Bảng 5.1. *Classification report trên tập Test – LR*

	Precision	**Recall**	**F1-score**	**Support**
1 – Tích cực	0.87	0.85	0.86	100
2 – Trung tính	0.76	0.70	0.73	53
3 – Tiêu cực	0.77	0.85	0.81	59
Trung bình	0.81	0.81	0.81	212

[13] What is difference between R2 score and Accuracy score: https://stackoverflow.com/questions/58163026/what-is-difference-between-metrics-r2-score-and-acccuracy-score



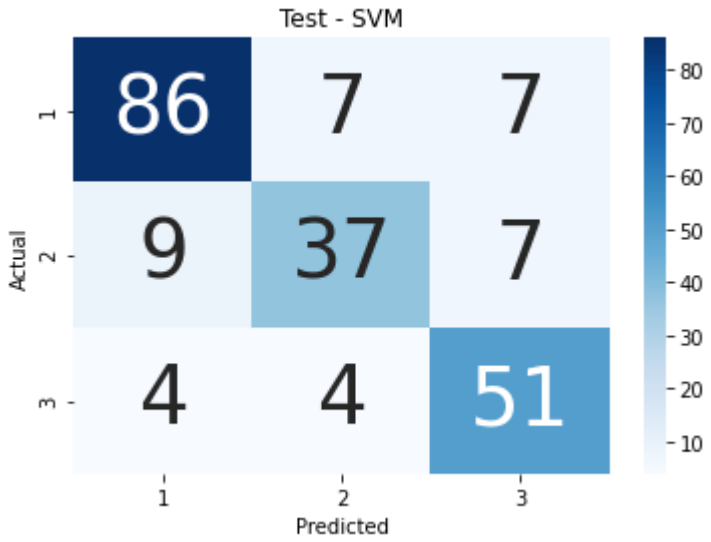
Hình 5.2. *Biểu đồ các giá trị độ đo Classification - LR trên tập Test*

Đối với mô hình hồi quy Logistic, ta nhận thấy:

- Kết quả đánh giá trung bình của mô hình phân loại là 81%
- Từ kết quả precision, recall và F1-score của mô hình sử dụng Logistic Regression được thể hiện trong Bảng và hình. Mô hình phân loại tương đối tốt đối với nhãn “1” (86%) và nhãn “3” (81%) trong khi đó với nhãn “2” (73%) kết quả không cao bằng.
- Điều này cũng có thể hiểu được do sự không cân đối của bộ dữ liệu, dữ liệu với nhãn “2” chỉ chiếm 25% trên bộ dữ liệu.

4.2. Suport Vector Machine

4.2.1. Confusion matrix và accuracy



Hình 5.3. *Confusion matrix trên tập Test - SVM*

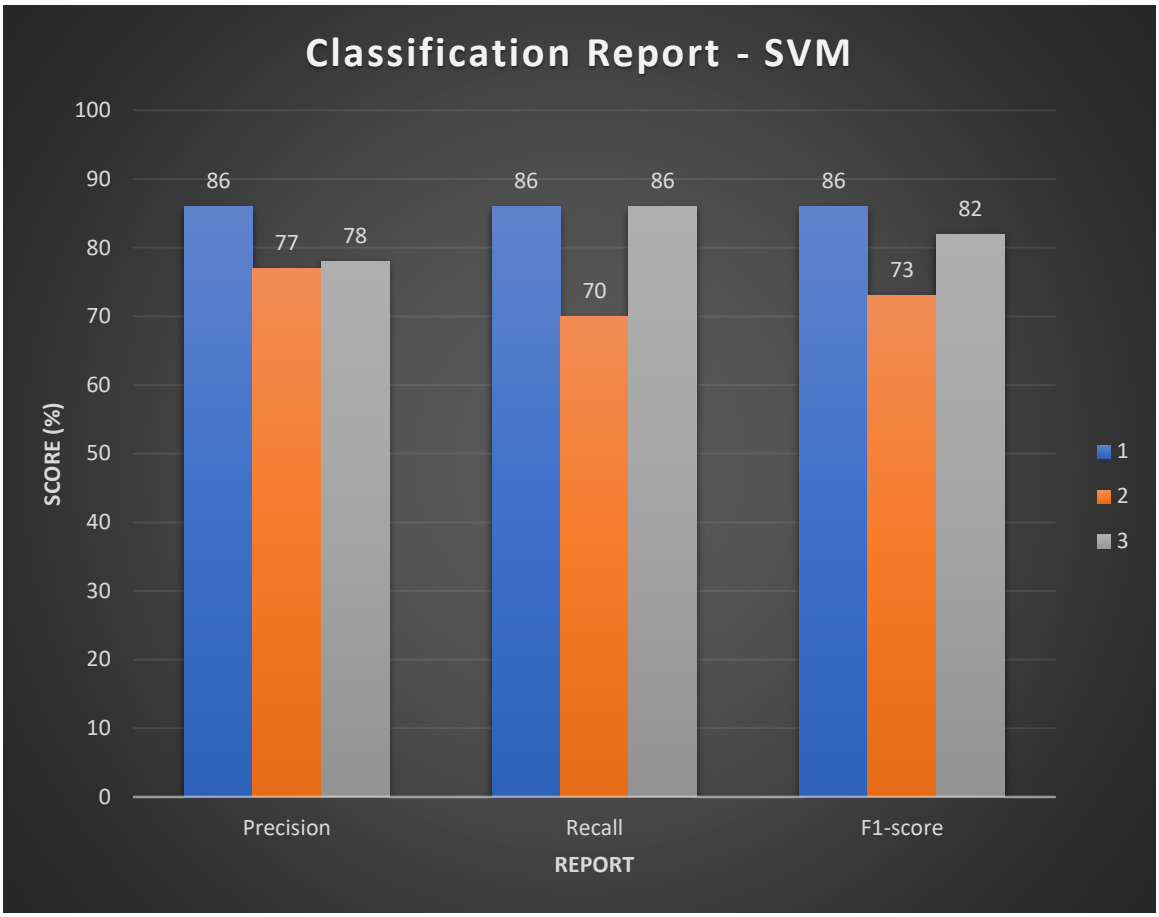
Theo ma trận trên, ta thấy:

- Số lượng dữ liệu được phân loại đúng là 86 + 37 + 51 = 174 điểm dữ liệu.
- Số lượng dữ liệu phân loại sai là 7 + 7 + 7 + 9 + 4 + 4 = 38 điểm dữ liệu.
- Tỉ lệ điểm dữ liệu phân loại đúng là 172/212 = 0.82075%

4.2.2. Classification report

Bảng 4.2. *Classification report trên tập Test – SVM*

	Precision	**Recall**	**F1-score**	**Support**
1 – Tích cực	0.86	0.86	0.86	100
2 – Trung tính	0.77	0.70	0.73	52
3 – Tiêu cực	0.78	0.86	0.82	59
Trung bình	0.82	0.82	0.82	212



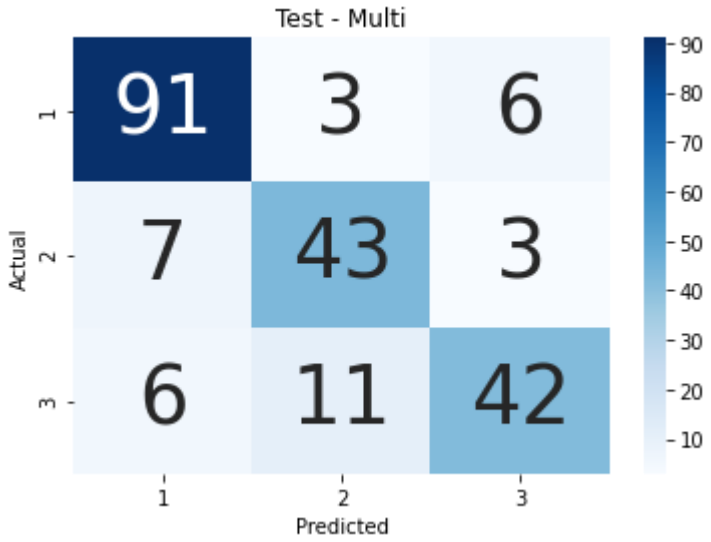
Hình 4.4. *Biểu đồ các giá trị độ đo Classification - SVM trên tập Test*

Đối với mô hình Support Vector Machine, ta nhận thấy:

- Kết quả đánh giá trung bình của mô hình phân loại là 82%
- Từ kết quả precision, recall và F1-score của mô hình sử dụng Support Vector Machine được thể hiện trong Bảng và hình. Mô hình phân loại tương đối tốt đối với nhãn "1" (86%) và nhãn "3" (82%) trong khi đó với nhãn "2" (73%) kết quả không cao bằng.
- Điều này cũng có thể hiểu được do sự không cân đối của bộ dữ liệu, dữ liệu với nhãn "2" chỉ chiếm 25% trên bộ dữ liệu.

4.3. MultinomialNB

4.3.1. Confusion matrix và accuracy



Hình 4.5. *Confusion matrix trên tập Test - MultinomialNB*

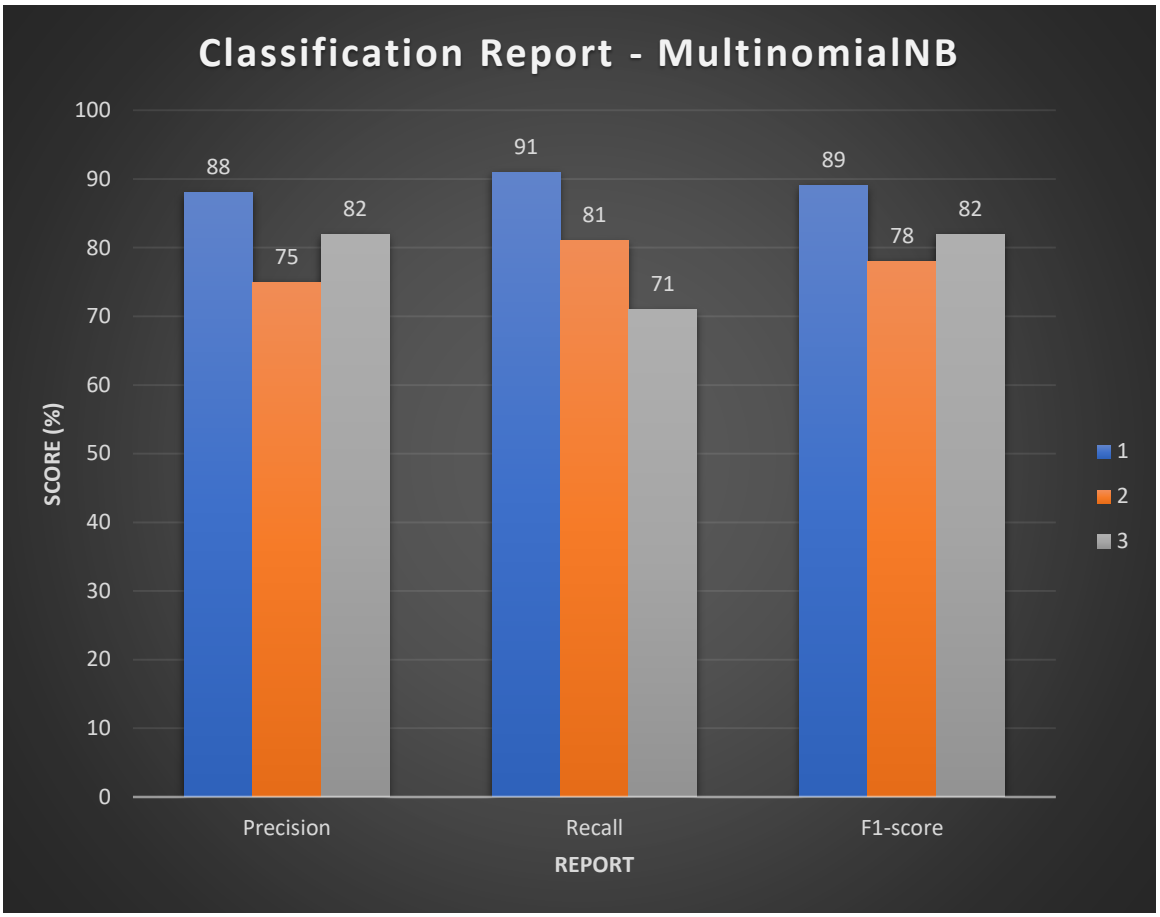
Theo ma trận trên, ta thấy:

- Số lượng dữ liệu được phân loại đúng là 91 + 43 + 42 = 176 điểm dữ liệu.
- Số lượng dữ liệu phân loại sai là 3 + 6 + 3 + 7 + 11 + 6 = 36 điểm dữ liệu.
- Tỉ lệ điểm dữ liệu phân loại đúng là 172/212 = 0.83018%

4.3.2. Classification report

Bảng 4.3. *Classification report trên tập Test – MultinomialNB*

	Precision	**Recall**	**F1-score**	**Support**
1 – Tích cực	0.88	0.91	0.89	100
2 – Trung tính	0.75	0.81	0.78	53
3 – Tiêu cực	0.82	0.71	0.76	59
Trung bình	0.83	0.83	0.83	212



Hình 4.6. *Biểu đồ các giá trị độ đo Classification - MultinomialNB trên tập Test*

Đối với mô hình MultinomialNB, ta nhận thấy:

- Kết quả đánh giá trung bình của mô hình phân loại là 81%
- Tương tự với 2 mô hình trên, các kết quả precision, recall và F1-score của mô hình sử dụng MultinomialNB được thể hiện trong Bảng và hình. Mô hình phân loại tương đối tốt đối với nhãn "1" (89%) và nhãn "3" (82%), tốt hơn so với nhãn "2" (78%).
- Điều này cũng có thể hiểu được do sự không cân đối của bộ dữ liệu, dữ liệu với nhãn "2" chỉ chiếm 25% trên bộ dữ liệu.

CHƯƠNG 5. KẾT LUẬN

Trong bài báo cáo này, chúng tôi đã trình bày một bộ dữ liệu phân loại cảm xúc của các bình luận về một sản phẩm âm nhạc, mà cụ thể ở đây là "Justin Beiber – What do you mean" bằng tiếng Anh. Chúng tôi đã tiến hành phân tích chi tiết những khó khăn liên quan đến việc thu thập và chú thích tập dữ liệu này. Chúng tôi đã thực hiện việc xây dựng các mô hình học máy để giải quyết bài toán phân loại trên và cho thấy mô hình MultinomialNB thực hiện tốt hơn so với các mô hình Logistic Regression và Support Vector Machine.

Nếu giả sử về tính độc lập được thoả mãn (dựa vào bản chất của dữ liệu), NBC được cho là cho kết quả tốt hơn so với SVM và Logistic Regression khi có ít dữ liệu training. [3]

Bảng 5.1. *Bảng so sánh kết quả độ đo các mô hình phân loại*

STT	Model + Features	Precision	Recall	F1-score	Accuracy
1	**LR + CountVectorizer**	0.81	0.81	0.81	0.81
2	**SVM + CountVectorizer**	0.82	0.82	0.82	0.82
3	**MultinomialNB + CountVectorizer**	0.83	0.83	0.83	0.83

Các mô hình dựa trên BOW trong hầu hết các nhiệm vụ phân loại đa nhãn. Các nhãn chú thích khác nhau và các nhóm tương đương sẽ giúp chúng tôi thực hiện việc học chuyển và điều tra cách thông tin đa phương thức trên các Youtube, dữ liệu không nhãn bổ sung, chuyển đổi nhãn và chia sẻ thông tin nhãn có thể tăng hiệu suất phân loại trong tương lai.

TÀI LIỆU THAM KHẢO

[1] H. D. Thoi, "Machine Learning cho người mới bắt đầu (Part 2)," VIBLO, 2018. [Online]. Available: https://viblo.asia/p/machine-learning-cho-nguoi-moi-bat-dau-part-2-naQZR1WXKvx.

[2] V. H. Tiệp, "Support Vector Machine," Machine Learning cơ bản, 2017. [Online]. Available: https://machinelearningcoban.com/2017/04/09/smv/.

[3] V. H. Tiệp, "Naive Bayes Classifier," Machine Learning cơ bản, 2017. [Online]. Available: https://machinelearningcoban.com/2017/08/08/nbc/.

[4] egbertbouman, "Youtube comment downloader," Github, 2020. [Online]. Available: https://github.com/egbertbouman/youtube-comment-downloader.

[5] D. K. 1, "Bag of words (BoW) model in NLP," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/bag-of-words-bow-model-in-nlp/.

[6] S.-l. developers, "sklearn.feature_extraction.text.CountVectorizer," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[7] GeeksforGeeks, "Confusion Matrix in Machine Learning," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/.

[8] Đ. A. Thi, "Hiểu confusion matrix," Math2it, 2019. [Online]. Available: https://math2it.com/hieu-confusion-matrix/.

[9] V. H. Tiệp, "Bài 15: Overfitting," Machine Learning cơ bản, 2017. [Online]. Available: https://machinelearningcoban.com/2017/03/04/overfitting/.

[10] "Regular expression operations," Python Software Foundation, 2020. [Online]. Available: https://docs.python.org/2/library/re.html.

[11] M. Younus, "Convert emoji unicode to TEXT in Python," Stack Overflow, 2018. [Online]. Available: https://stackoverflow.com/questions/47489836/convert-emoji-unicode-to-text-in-python.

[12] S.-l. developers, "Cross-validation: evaluating estimator performance," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html.

[13] S.-l. developers, "sklearn.metrics.r2_score," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

[14] S.-l. developers, "sklearn.metrics.classification_report," Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html.

[15] PV8, "What is difference between metrics.r2_score and acccuracy_score," Stack Overflow, 2019. [Online]. Available: https://stackoverflow.com/questions/58163026/what-is-difference-between-metrics-r2-score-and-acccuracy-score.

PHỤ LỤC

Source code của đồ án được lưu trên github theo link dưới.

Link github: https://github.com/18520426/UIT-DS102-Final-Report