# Phân Loại Cảm Xúc Cho Tiếng Việt Với Các Bình Luận Về Son

Nguyễn Trọng Ân*, Dương Văn Bình*, Hà Như Chiến*,
Dương Thị Hồng Hạnh*, Trần Thị Mỹ Linh*, Nguyễn Văn Kiệt†
Email: *{18520434, 18520505, 18520527, 18520711, 18520999}@gm.uit.edu.vn
†{kietnv}@uit.edu.vn

***Tóm tắt nội dung*—Trong đồ án này, chúng tôi tiến hành xây dựng, phát triển và đánh giá bộ dữ liệu bình luận Tiếng Việt UIT - VLFC (Vietnamese Lipstick Feedbacks Corpus) từ các trang thương mại điện tử hàng đầu tại Việt Nam. Chúng tôi tiến hành thu thập, xử lý, tạo bộ dữ liệu, sau đó tiến hành thực nghiệm, đánh giá bằng các phương pháp học máy truyền thống mà ở đây là các mô hình Naive Bayes, Logistic Regression đồng thời áp dụng các kỹ thuật phân loại văn bản state-of-the-art (SOTA) như Bi-LSTM, Fully Connected Network và Ensample dùng cho các bài toán phân loại văn bản Tiếng Việt trên bộ dữ liệu đã xây dựng. Kết quả thu được tương đối tốt với các thuật toán như SVM hay Logistic Regression có kết quả các độ đo F1-score lần lượt là 0.6951, 0.6978 và các mô hình SOTA cho kết quả độ đo F1-score là 0.7390, 0.7028, 0,7452 trên tập kiểm thử (test set). Từ đó, chỉ ra mô hình tối ưu nhất có thể, phân tích ưu điểm và nhược điểm của các mô hình đối với bài toán phân loại này, đề xuất các giải pháp để cải thiện chất lượng bộ dữ liệu, nâng cao hiệu suất phân loại của các mô hình trong tương lai.**

***Từ khóa*—**Xây dựng bộ dữ liệu, Phân loại văn bản Tiếng Việt, Phân tích cảm xúc

## I. Giới thiệu

Công nghệ thông tin ngày càng phát triển, hiện đại hơn nên sàn giao dịch điện tử không phải là thuật ngữ quá xa lạ, đó là nơi người mua và người bán có thể trao đổi, mua - bán các sản phẩm, dịch vụ một cách trực tuyến trên một website. Nó tồn tại cùng với thực tiễn, mang lại rất nhiều hữu ích cho người dùng, trở thành môi trường kinh doanh không thể thiếu không chỉ với người tiêu dùng mà còn với các thương nhân buôn bán. Tuy nhiên, một điều mà thương mại điện tử thiếu và thứ mà có lẽ nó sẽ thiếu mãi mãi là thiếu cảm giác vật lý. Hình ảnh, âm thanh và video là một chuyện, nhưng giữ vật phẩm bằng mười ngón tay của mình là một điều hoàn toàn khác. Đó là lí do vì sao đối với người tiêu dùng, các bình luận của người mua trước rất quan trọng - chất lượng sản phẩm ra sao, hình thức như thế nào sẽ được thể hiện qua những bình luận này. Mỗi bình luận sẽ là một trải nghiệm sản phẩm thực tế, nó mang một cảm xúc riêng của khách hàng. Với hàng nghìn sản phẩm trên sàn thương mại điện tử, mỗi sản phẩm lại có hàng trăm hàng nghìn lời bình luận, việc phân loại cảm xúc của mỗi bình luận bằng cách thủ công là một công việc không hề dễ dàng. Thay vào đó, việc nhận diện cảm xúc của các bình luận về sản phẩm một cách tự động giúp cho việc đánh giá trở nên khách quan và hiệu quả hơn.

Ở đây chúng tôi lựa chọn phân loại cảm xúc bình luận của sản phẩm son trên các trang thương mại điện tử lớn theo các thái độ tích cực, trung tính, tiêu cực. Mục tiêu đặt ra bài toán này là xây dựng và phát triển bộ dữ liệu bình luận về sản phẩm son trên các trang thương mại điện tử lớn tại Việt Nam. Từ đó, áp dụng và đánh giá hiệu suất các mô hình học máy, học sâu để tiến hành phân loại bình luận.

Trong bài báo cáo này, chúng tôi tập trung vào giới thiệu các thông tin liên quan đến bài toán xây dựng bộ dữ liệu bình luận về sản phẩm son. Trong mục II , chúng tôi sẽ trình bày một số công trình nghiên cứu liên quan. Tiếp theo ở mục III, chúng tôi trình bày chi tiết về quá trình thu thập, xử lý và tạo ra bộ dữ liệu. Trong mục IV, các giải pháp, mô hình được chúng tôi trình bày và đồng thời, kết quả thử nghiệm sẽ được đánh giá, phân tích ở mục V. Cuối cùng, mục VI sẽ là kết luận và hướng phát triển trong tương lai cho các bài toán phân loại cảm xúc bình luận của sản phẩm son trên các trang thương mại điện tử lớn.

## II. Các công trình nghiên cứu liên quan

Binh Thanh Kieu và cộng sự (2010)[1] nghiên cứu phân tích cảm xúc cho Tiếng Việt ở mức độ câu văn và xây dựng hệ thống cơ sở cho phương pháp này dựa trên công cụ hỗ trợ là GATE framework. Các kết quả thử nghiệm của nhóm tác giả được thực hiện trên bộ dữ liệu về các bình luận về sản phẩm là máy tính. Theo nhóm tác giả đây là công trình đầu tiên nghiên cứu về cảm xúc cho Tiếng Việt ở cấp độ câu văn. Bộ dữ liệu được thu thập từ nguồn dữ liệu trực tuyến trên các trang quảng cáo thương mại điện tử gồm 3971 câu trong 20 tập tài liệu tương ứng với 20 sản phẩm. Hướng tiếp cận của nhóm tác giả ở cấp độ từ là gắn thẻ PosWord (từ tích cực) và NegWord (từ tiêu cực), ở mức độ câu là gắn thẻ PosSen (câu tích cực) và NegSen (câu tiêu cực), và MixSen (câu hỗn hợp) để phân biệt câu tích cực và câu tiêu cực với câu vừa tích cực vừa tiêu cực. Độ chính xác trên tập train (3182 bình luận) và test (789 bình luận) cho với 3 loại câu PosSen, NegSen và MixedSen lần lượt là 67.35% và 61.16%.

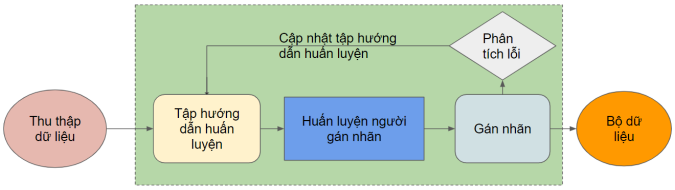
Son Trinh và cộng sự (2016)[2] đã nghiên cứu phân tích cảm xúc cho Tiếng Việt từ các bình luận trên Mạng xã hội Facebook. Hướng tiếp cận của nhóm tác giả là việc xây dựng một tập từ điển cảm xúc cơ sở phục vụ việc phân tích cảm xúc cho Tiếng Việt. Bộ từ điển được xây dựng gồm có các phần như: từ điển cho danh từ, động từ, tính từ, và trạng từ. Nhóm tác giả có còn đề xuất thêm các đặc trưng dựa trên cơ sở phân tích cảm xúc trên Tiếng Anh với một số điều chỉnh để phù hợp với đặc trưng của Tiếng Việt và sử dụng SVM classifier để xác định cảm xúc bình luận người dùng. Số lượng từ trong các tập từ điển của danh từ, động từ, tính từ và trạng từ lần lượt là 1546, 1108, 2357, 749 và mỗi từ là một cặp gồm một từ và một số nguyên từ -5 (rất tiêu cực) tới 5 (rất tích cực), không có từ nào đi cùng với giá trị 0. Ngoài ra còn có từ điển cho những từ tăng cường phổ biến trong Tiếng Việt, mỗi cặp sẽ là một từ và một số thực từ để chỉ mức độ ('ít', -1.5), ('chút ít', -0.9). Tập dữ liệu dùng để thử nghiệm được thu thập với 3 chủ đề chính là Giáo dục, Phim, Thể thao với tổng là 885 bình luận. Bộ dữ liệu cũng được gán nhãn bằng tay. Kết quả đạt được sau khi thử nghiệm trên mô hình thì độ chính xác đạt được 89.8%.

## III. Bộ dữ liệu

Trong phần này, chúng tôi trình bày các thông tin cơ bản về bộ dữ liệu, quy trình thu thập và các thách thức mà chúng tôi phải đối mặt trên bộ dữ liệu UIT- VLFC.

### A. Thu thập và gán nhãn dữ liệu

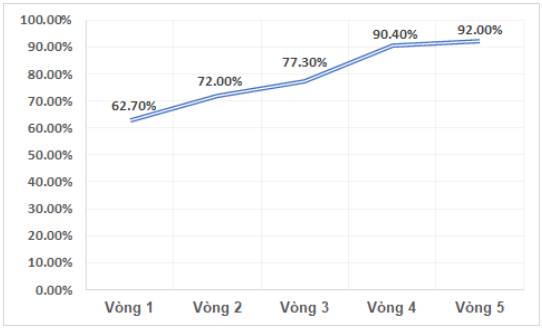
Chúng tôi tiến hành thu thập dữ liệu dựa trên 3 trang thương mại điện tử uy tín hàng đầu Việt Nam là: Tiki, Lazada, Shoppe. Dữ liệu được thu thập tự động theo các mẫu son được bán trên các trang bằng cách sử dụng thư viện Beautiful Sop [1] trên ngôn ngữ Python. Sau đó chúng tôi tiến hành gán nhãn dựa trên Bộ hướng dẫn gán nhãn (guideline) được xây dựng và cập nhật nhiều lần song song với việc gán nhãn. Cuối cùng, chúng tôi đã xây dựng được bộ dữ liệu với 12228 điểm dữ liệu.




Hình 1: Quy trình xây dựng bộ dữ liệu UIT -VLFC

### B. Quy trình gán nhãn

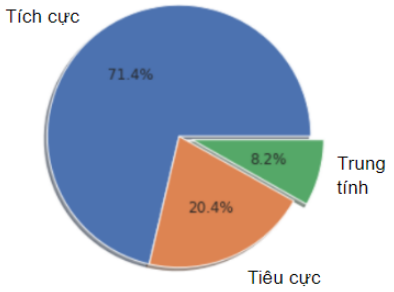
Trước khi tiến hành gán nhãn chính thức, chúng tôi thực hiện 5 vòng gán nhãn chung để rút luật và huấn luyện người gán nhãn. Mỗi vòng gán nhãn 100 câu, ghi lại độ đồng thuận (được tính theo Fleiss'Kappa[3]), kiểm tra những câu bất đồng để rút luật, rồi dùng những luật vừa có để gán nhãn ở vòng tiếp theo. Thực hiện xong 5 vòng, độ đồng thuận chúng tôi đạt được là 92.0% ổn định so với vòng 4 là 90.4%. Hình 2 cho thấy sự cải thiện độ đồng thuận qua các lần huấn luyện. Đối với các dòng dữ liệu còn lại, các thành viên tự gán riêng lẻ, câu nào nằm ngoài những luật đã có thì ghi chú lại sau đó những người gán nhãn sẽ thảo luận để đưa ra nhãn cuối cùng.




Hình 2: Độ đồng thuận qua các lần gán nhãn

### C. Thông tin bộ dữ liệu

Bộ dữ liệu hoàn chỉnh bao gồm 12,228 điểm dữ liệu với 2 thuộc tính: "comments" (nội dung bình luận của khách hàng đối với sản phẩm và dịch vụ được cung cấp bởi cửa hàng), "label" (thái độ đánh giá của khách hàng thể hiện thông qua bình luận). Trong đó, thuộc tính "label" là thuộc tính mục tiêu cần dự đoán, gồm 3 loại thái độ tương ứng với 3 loại nhãn có tỉ lệ các nhãn như sau:




Hình 3: Tỷ lệ phần trăm các nhãn

Ví dụ mẫu được thể hiện cụ thể thông qua Bảng I và Bảng II bên dưới.

Bảng I: Các loại nhãn và thái độ tương ứng

| label | Thái độ tương ứng |
|---|---|
| 0 | Tích cực (Possitive) |
| 1 | Tiêu cực (Negative) |
| 2 | Trung tính (Neutral) |

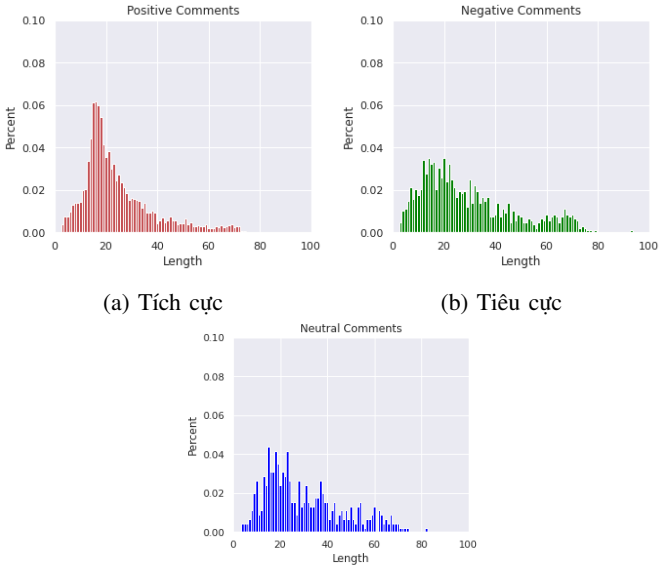
Bảng II: Một số dữ liệu mẫu theo từng loại nhãn

| STT | comments | label |
|---|---|---|
| 1 | Sản phẩm cực tốt đáng đồng tiền nói chung là rất oke nha cả nhà đó là màu môi em đánh nhe yêu shop quá mua hai cây được gói trong cái hộp cute | 0 |
| 2 | Chất son lên môi rất mướt lên mau rất đẹp cực ưng :)) | |
| 3 | son đẹp cực lun ạ | |
| 4 | Giao hàng hơi lâu nhưng son đẹp | |
| 5 | :(( son dính, có bóng nhưng ít, có mùi nước rửa bát | 1 |
| 6 | Màu lên đẹp nhưng chất son không thích. Mùi quá ngọt. Lúc mua khuyến mãi có kèm túi nhưng không có. Nhưng thôi kệ không sao.Nói chung là son không thích ạ. | |
| 7 | Đặt 1 mã gửi 1 mã khác, nhắn tin hỏi cửa hàng thì hoạt động trực tuyến mà không thèm trả lời. Cho 1 sao vì ghét cái thái độ bán hàng. | |
| 8 | Ban đầu nghi, màu sẽ nghiêng về dạng cam cháy như ver3 màu 12.Tuy nhiên, lúc thử giống màu cam neon hơn, với mình đánh đậm lên màu đẹp sẽ đẹp hơn | |
| 9 | Bình thường | 2 |
| 10 | Cảm giác chất son hơi lỏng | |
| 11 | chất son dễ tán. màu lên đẹp. tuy nhiên rất lem cốc, không khô hẳn. | |
| 12 | Cho 3 sao thôi vì thông tin cung cấp chưa đầy đủ về deal | |

[1]Beautiful Soup - https://www.crummy.com/software/BeautifulSoup/bs4/doc/up

### D. Thăm dò dữ liệu

- Bình luận có độ dài dài nhất: 108 tiếng.
- Bình luận có độ dài ngắn nhất: 1 tiếng.
- Độ dài các bình luận của khách hàng theo từng loại nhãn được mô tả bởi Hình 4 :

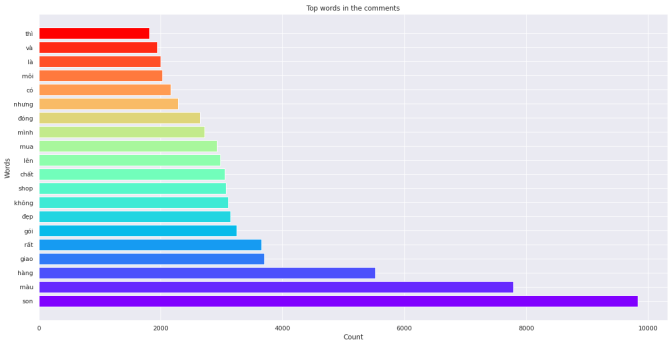



(a) Tích cực (b) Tiêu cực

(c) Trung tính

Hình 4: Độ dài bình luận của từng loại nhãn.

  - Hình dạng biểu độ thống kê số lượng bình luận theo độ dài của nhãn tích cực khác hoàn toàn so với nhãn tiêu cực và trung tính. Phần lớn bình luận tích cực có độ dài phổ biến trong khoảng tù 15-15 tiếng.
  - Những bình luận dài thường là bình luận trung tính và tiêu cực.
- Các từ xuất hiện nhiều trong bình luận của khách hàng được biểu diễn bởi Hình 5 :




Hình 5: Top 20 từ phổ biến nhất trong bộ dữ liệu UIT -VLFC.

- Wordcloud theo từng loại nhãn biểu diễn ở Hình 6 :


(a) Tích cực (b) Tiêu cực

(c) Trung tính

Hình 6: Wordcloud của từng loại nhãn.

  - Những bình luận tích cực đa số đề cập nhiều đến dịch vụ như đóng gói, giao hàng.
  - Bình luận tiêu cực nói nhiều về son hơn so với dịch giao hàng. Xuất hiện nhiều từ thể hiện rõ thái độ tiêu cực như 'thất vọng', do đó từ 'thất vọng' sẽ dễ bị thiên vị trong lúc dự đoán kể cả khi nó đi kèm với các từ hơi,không.
  - Bình luận trung tính chứa cả những từ xuất hiện nhiều ở bình luận tích cực lẫn tiêu cực nên dễ bị nhầm lẫn trong quá trình dự đoán.

### E. Thách thức của bộ dữ liệu

Trong quá trình xây dựng bộ dữ liệu UIT - VLFV, chúng tôi đã gặp phải một số thách thức:

- Dữ liệu mất cân bằng: Kết quả gán nhãn cho thấy dữ liệu bị mất cân bằng, nhãn tích cực quá nhiều 71.4% nhãn tiêu cực và trung tính chỉ chiếm 26.6%
- Ngôn ngữ nhập nhằng gây khó khăn trong quá trình gán nhãn dữ liệu dẫn đến việc tốn thời gian trong công đoạn xây dựng bộ luật gán nhãn và huấn luyện người gán nhãn.

## IV. Phương pháp tiếp cận

### A. Tiền xử lý dữ liệu

Sau khi có bộ dữ liệu được gán nhãn đầy đủ như đã đề cập ở Phần III, để xây dựng mô hình trên bộ dữ liệu trước tiên chúng ta cần làm sạch bộ dữ liệu:

- Loại bỏ các dấu câu
  Ví dụ: "son đẹp, lâu trôi :)) → "son đẹp lâu trôi".
- Đưa về văn bản viết thường
  Ví dụ: "Chất son mịN đẹp" → "chất son mịn đẹp".
- Xử lý các chữ kéo dài, loại bỏ các chữ không có ý nghĩa và một số chữ viết tắt
  Ví dụ: "ko đc tốttttttt gojoejkksie" → "không được tốt"
- Xử lý emoji: gộp các emoji giống nhau thành một emoji và chuyển emoji thành chữ viết
  Ví dụ: "yêu son nhiều ☺ ☺ hi ☺ " → "yêu son nhiều hi smile_face"
- Thực hiện tách từ tiếng Việt (sử dụng thư viện tách từ vncorenlp [2])
  Ví dụ: "giao hàng nhanh màu son ưng ý" → ['giao', 'hàng', 'nhanh', 'màu_son', 'ưng_ý']

Sau khi toàn bộ tập dữ liệu đã được làm sạch. Chúng tôi tiến hành chia bộ dữ liệu thành 3 tập train, val và test theo tỷ lệ 8:1:1. Tập

[2]VnCoreNLP - https://github.com/vncorenlp/VnCoreNLP

train sẽ được dùng để huấn luyện các mô hình, 2 tập val và test sẽ được dùng để đánh giá mô hình.

### B. Vector biểu diễn từ

Để có thể tiến hành huấn luyện mô hình thì các điểm dữ liệu phải được biểu diễn dưới dạng số học để máy có thể hiểu. Trong bài báo này, chúng tôi sử dụng CountVectorizer[3] để tạo các vector biểu diễn theo bộ từ điển được tạo ra sau bước tách từ.

### C. Mô hình

*1) Multinomial Naive Bayes:* Mô hình Naive Bayes một mô hình phân loại theo xác suất, với mỗi tập tài liệu d, các điểm dữ liệu chưa được gán nhãn sẽ được mô hình trả về một nhãn c tương ứng có xác suất hậu tiên nghiệm (posterior probability) cao nhất. Ý tưởng được biết đến từ công trình nghiên cứu của Bayes (1763)[4] và lần đầu được áp dụng cho phân loại chữ viết bởi F. Mosteller và D. L. Wallace(1964)[5].

Mô hình Multinomial Naive Bayes[4] chủ yếu được sử dụng trong phân loại văn bản nơi mà các vector đặc trưng (feature vectors) được tính bằng túi từ (Bag of words). Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài L 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.

*2) Random Forests[6]:* Mô hình Random Forests là mô hình phân loại phát triển dựa trên thuật toán cây quyết định (Decision Tree) với ý tưởng nhiều cây sẽ tốt hơn một cây đơn lẻ. Quy trình phân loại nhãn của mô hình gồm 2 giai đoạn: giai đoạn thứ nhất là lấy kết quả dự đoán nhãn của K cây quyết định, giai đoạn thứ hai là tổng hợp (ensemble) để lựa chọn nhãn phổ biến nhất (majority voting) chính là nhãn cuối cùng mà mô hình trả về.

*3) Logistic Regression[7]:* là một thuật toán phân loại phổ biến khi nhắc tới các bài toán phân loại. Thuật toán dựa vào hàm Sigmoid để dự đoán nhãn của điểm dữ liệu với ngưỡng là 0.5.

*4) Support Vector Machine[8]:* SVM là thuật toán phổ biến trong Học máy. SVM được ứng dụng trong rất nhiều lĩnh vực với cả bài toán Hồi quy và Phân loại. SVM có thời gian huấn luyện khá tốt trên bộ dữ liệu lớn.

*5) Fully connected network:* Chúng tôi sử dụng 5 lớp Dense với Dropout sau mỗi lớp để đảm bảo cho việc tránh quá khớp.

*6) Bidirectional-LSTM:* LSTM[9] ra đời giúp khắc phục những hạn chế của các thuật toán trước đó. Ý tưởng của LSTM là kết nối các thông tin phía trước để dự đoán thông tin hiện tại vì vậy với các dữ liệu dạng chuỗi như các bình luận, ta sẽ giữ được ngữ nghĩa của câu giúp cho việc dự đoán trên dữ liệu mới đạt kết quả tốt hơn vì ý nghĩa của các từ hay xuất hiện cùng nhau sẽ được mô hình ghi nhớ. Bi-LSTM là phiên bản mở rộng của mô hình sử dụng LSTM với mục đích cải thiện hiệu suất của mô hình. Ý tưởng của Bi-LSTM là sẽ chạy LSTM hai lần trên chuỗi đầu vào (một lượt đi và một lượt về) việc này giúp cung cấp thêm thông tin về ngữ nghĩa cho mô hình, nhờ đó mà hiệu suất của mô hình được tăng lên.

### D. Ensemble Learning

Ensemble Systems[10][11] là một kỹ thuật giúp tăng cường độ chính xác cho các bài toán phân loại. Được sử dụng để giải quyết những vấn đề của Học máy như: trích chọn đặc trưng, ước lượng khoảng tin cậy, **mất cân bằng nhãn**,... . Có nhiều phương pháp ensemble khác nhau được phát triển để đáp ứng những mục đích sử dụng của nhiều bài toán khác nhau, trong đồ án này chúng tôi sử dụng Emsemble Majority Voting[12].

Trong đồ án này, chúng tôi sử dụng nhiều mô hình khác nhau để đánh giá bộ dữ liệu vì vậy việc sử dụng phương pháp này giúp tối ưu kết quả của các mô hình và khắc phục việc mất cân bằng trong dự đoán các nhãn vì bộ dữ liệu mất cân bằng.

## V. Thử nghiệm và thảo luận

### A. Thông số mô hình

Tất cả các phương pháp đều được so sánh trên bộ dữ liệu đã được phân chia giống nhau và kết quả được báo cáo dựa trên tập validation và tập kiểm thử. Tại đây, chúng tôi tiến hành tinh chỉnh các thông số để huấn luyện cho cả 6 mô hình được đưa ra là: MultinomialNB, Random Forest, Logistic Regression, SVM, Fully Connected Net và Ensample để có thể so sánh và đánh giá đa chiều hiệu suất của các mô hình trên.

- **Multinomial Naive Bayes**: Chúng tôi sử dụng MutinomialNB với alpha = 1.0.
- **Logistic Regression (LR)**: Cài đặt các tham số mô hình Logistic Regression với C = 1.0, class_weight='balanced', max_iter = 100, multi_class = 'auto', penalty = 'l2', solver = 'lbfgs', warm_start = True.
- **Random Forest**: Cài đặt các tham số mô hình Random Forest với bootstrap=True, class_weight='balanced', criterion = 'gini', max_features = 'auto', min_samples_leaf = 1, min_samples_split = 2, n_estimators = 100, warm_start = False.
- **Support Vector Machine**: Cài đặt các tham số mô hình Support Vector Machine với C = 1.0, cache_size = 200, class_weight='balanced', decision_function_shape = 'ovo', degree = 3, gamma = 'scale', kernel='rbf', max_iter = -1.
- **Fully Connected Neural Network**: Mạng neural gồm 5 lớp Dense, trong đó 4 lớp Dense đầu tiên có số nốt lần lượt là 256, 64, 32, 16, áp dụng hàm kích hoạt 'relu'. Riêng lớp Dense cuối cùng gồm 3 nốt theo sau bởi hàm kích hoạt 'softmax' vì mục tiêu là phân loại 3 nhãn. Dropout phía sau các lớp Dense có rate lần lượt là 0.1, 0.2, 0.2, 0.3 để đảm bảo mô hình không đi vào quá khớp.
- **Bi-LSTM:** Sử dụng framework Keras để cài đặt gồm 1 lớp Embedding, 1 lớp Bidirectional với số nốt 50, 1 lớp MaxPooling + Dropout với rate=0.1, 1 lớp Dense với số nốt 50 + Dropout với rate=0.1, cuối cùng là một lớp Dense dùng để phân loại nhãn với 3 nốt. Các hàm kích hoạt được sử dụng cho 2 lớp Dense lần lượt là 'relu' và 'softmax'. Với lớp Embedding chúng tôi có tham khảo sử dụng Vector embedding FastText[5] cho Tiếng Việt.

### B. Phân tích kết quả và thảo luận

Trong phần này, chúng tôi trình bày kết quả thử nghiệm đạt được với các mô hình đã đề cập bên trên. Kết quả phân loại của mô hình phân loại được đánh giá dựa trên thang đo F1-score. Bảng III dưới đây mô tả độ chính xác của 7 mô hình trên bộ dữ liệu UIT-VLSP. Trong đó, mô hình Multinomial Naïve Bayes là mô hình đánh giá đầu tiên cho việc cài đặt dễ dàng và triển khai nhanh. Đại diện cho mạng Neural chúng tôi có tiến hành cài đặt mô hình Bi-LSTM có nhiều đặc điểm phù hợp với dữ liệu dạng chuỗi. Chúng tôi tiến hành tổng hợp dự đoán của tất cả các mô hình đã triển khai được và nhận được kết quả tốt hơn so với kết quả đơn mô hình.

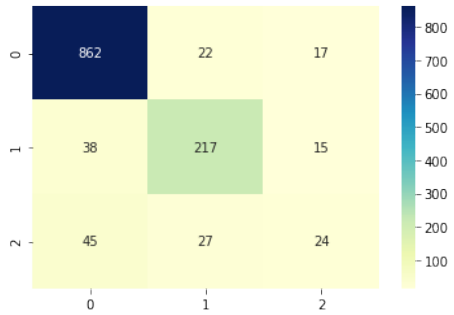
[3]https://scikit-learn.org/stable/modules/feature_extraction.html

[4]Multinomial Naive Bayes - https://web.stanford.edu/ jurafsky/slp3/4.pdf

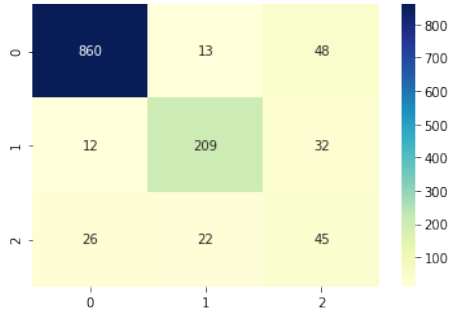
[5]VietNLP - https://github.com/vietnlp

Bảng III: Kết quả đánh giá trên bộ dữ liệu UIT -VLFC

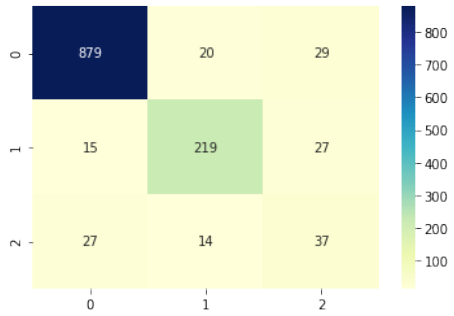
| Mô hình | Val F1-Score | Test F1–Score |
|---|---|---|
| **Multinomial Naïve Bayes** | 0.6865 | 0.6363 |
| **Random Forest** | 0.6702 | 0.6595 |
| **Logistic Regression** | 0.7195 | 0.6951 |
| **SVM** | 0.7197 | 0.6978 |
| **Bi-LSTM** | 0.7417 | 0.7390 |
| **Fully Connected Net** | 0.7571 | 0.7028 |
| **Ensemble** | **0.7615** | **0.7452** |




Hình 7: Confusion Matrix của mô hình Naive Bayes




Hình 8: Confusion Matrix của mô hình Bi-LSTM




Hình 9: Confusion Matrix của phương pháp Ensemble

Từ các kết quả ở Hình 7, 8 và 9, chúng tôi tiến hành so sánh khả năng dự đoán nhãn chính xác của 4 mô hình nêu trên:

Chúng ta có thể nhận thấy không có sự khác biệt quá lớn về độ chính xác trong việc phân loại các mô hình. Các mô hình truyền thống như Logistic Regression, SVM cho kết quả tương đối tốt, cùng với thời gian huấn luyện nhanh, dễ cài đặt, không yêu cầu cấu hình mạnh mẽ. Đây chính là các mô hình nên được cân nhắc khi cần giải quyết các bài toán phân loại, mà cụ thể ở đây là phân loại văn bản. Bên cạnh đó, mô hình Bi-LSTM hứa hẹn sẽ mang lại những bước chuyển biến lớn trong các tác vụ phân loại văn bản với các mô hình hiệu suất cao, hiệu quả. Đồng thời bằng phương pháp Ensemble dựa trên kết quả của các mô hình cũng cho kết quả rất tốt.

### *C. Phân tích lỗi*

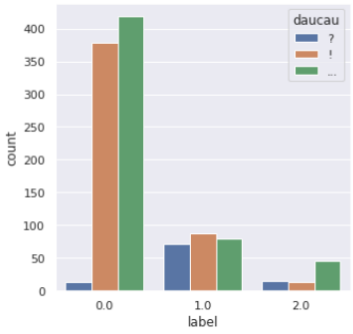
Trong bộ dữ liệu, vẫn còn một số bình luận bị phân loại nhầm, phần lớn nguyên nhân là do sự phức tạp trong ngữ nghĩa của Tiếng Việt. Bảng IV trình bày các ví dụ về các lỗi phân loại. Chúng ta có thể thấy rằng, các bình luận bị phân loại nhầm do có quá nhiều ý, mức độ khác biệt giữa tích cực và tiêu cực không rõ ràng; những từ "hơi", "nhưng", "rất" làm tăng hoặc giảm mức độ sắc thái của câu; nhiều bình luận không viết dấu dẫn đến sai lệch ý nghĩa của từ; ngoài ra còn có những câu chứa yếu tố phủ định bị nhầm thành tiêu cực.

Bảng IV: Một số ví dụ về lỗi phân loại trên bộ dữ liệu UIT - VLFC

| Bình luận | Thực tế | Dự đoán |
|---|---|---|
| Giao đến thế này đây buồn hết biết nhưng chất son cũng khá là oke son không bị bể vỡ ở đâu cả nên đây là lỗi của chính son chứ kh phải bên vận chuyển. Mùi thơm nhưng son khá đậm | Trung tính | Tiêu cực |
| Giao hàng nhanh , đóng gói chắc chắn :))) màu cũng khá bềnh, mùi tạm được | Trung tính | Tích cực |
| Son màu 23, 25 dùng k cần make up vẫn rất xinh, nhưng son 23 mình thấy bị hơi dừ khi sử dụng với mình | Trung tính | Tích cực |
| Hơi lâu 1 xíu nhưng nhìn vỏ thôi là ưng lắm ạ!!! | Tích cực | Trung tính |
| S lan son k dc min nhu lan trc | Tiêu cực | Tích cực |
| sản phẩm đóng gói đẹp cẩn thận nhưng son mình không ưng lắm vì không hề lì và rất nhanh trôi son không bám môi lâu lên tầm 2 tiếng là bay hết | Tiêu cực | Trung tính |
| Màu ưng ! Son lâu trôi ! Giá hợp lý ! Không thất vọng ! | Tích cực | Tiêu cực |

## VI. Kết luận và hướng phát triển

- Từ kết quả thực nghiệm được trình bày trong đồ án, ta nhận thấy các mô hình truyền thống đem lại kết quả khá tốt, và việc kết hợp các mô hình mang lại hiệu quả tốt nhất trên tập dữ liệu này cho đến hiện tại. Trong tương lai, chúng tôi dự định sẽ kết hợp thêm nhiều mô hình để nâng cao hiệu suất. Mô hình Bi-LSTM cũng cho thấy được các mô hình Deep Learning có tiềm năng phát triển, trong tương lai chúng tôi sẽ tiến hành thử nghiệm trên nhiều mô hình khác như PhoBERT, XLM-R,...
- Ngoài ra, chúng tôi cũng nhận thấy được một số bất lợi của bộ dữ liệu, cụ thể là việc gán nhãn còn khá phức tạp, cũng như chưa phân chia được từng vấn đề cần quan tâm (dịch vụ, sản phẩm, giá cả,...) nên chúng tôi sẽ cố gắng xây dựng lại bộ dữ liệu tốt hơn.
- Cuối cùng, trong quá trình thăm dò dữ liệu, chúng tôi cũng phát hiện được một điểm thú vị về mối liên hệ giữa dấu câu với thái độ của bình luận, cụ thể trong hình 10, và sẽ tiếp tục nghiên cứu về vấn đề này.




Hình 10: Biểu đồ thống kê số lượng xuất hiện của dấu câu theo từng loại nhãn

## TÀI LIỆU


[1] B. T. Kieu and S. B. Pham, "Sentiment analysis for vietnamese," in *2010 Second International Conference on Knowledge and Systems Engineering*, pp. 152–157, IEEE, 2010.

[2] S. Trinh, L. Nguyen, M. Vo, and P. Do, "Lexicon-based sentiment analysis of facebook comments in vietnamese language," in *Recent developments in intelligent information and database systems*, pp. 263–276, Springer, 2016.

[3] D. G. Seigel, M. J. Podgo, and N. A. Remaley, "Acceptable values of kappa for comparison of two groups," *American journal of epidemiology*, vol. 135, no. 5, pp. 571–578, 1992.

[4] T. Bayes, "Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s," *Philosophical transactions of the Royal Society of London*, no. 53, pp. 370–418, 1763.

[5] F. Mosteller and D. L. Wallace, "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963.

[6] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Critical care*, vol. 9, no. 1, p. 112, 2005.

[8] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] R. Polikar, "Ensemble learning," in *Ensemble machine learning*, pp. 1–34, Springer, 2012.

[11] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.

[12] A. K. Chowdhury, D. Tjondronegoro, V. Chandran, and S. G. Trost, "Ensemble methods for classification of physical activities from wrist accelerometry," *Medicine & Science in Sports & Exercise*, vol. 49, no. 9, pp. 1965–1973, 2017.