IMDB Movie Reviews Sentiment Analysis

Ngô Gia Lâm
Khoa Khoa học và Kỹ thuật Thông Tin
Trường Đại học Công nghệ thông tin - VNUHCM
Thành Phố Hồ Chí Minh, Việt Nam
21521054@gm.uit.edu.vn

Phạm Lê Thành Phát
Khoa Khoa học và Kỹ thuật Thông Tin
Trường Đại học Công nghệ thông tin - VNUHCM
Thành Phố Hồ Chí Minh, Việt Nam
21521262@gm.uit.edu.vn

***Tóm tắt*—Trong đề tài này, chúng tôi tập trung vào bài toán phân tích cảm xúc dựa trên tập dữ liệu bình luận phim ảnh trên trang IMDB được công bố tại tạp chí ACL 2011 [1]. Mục tiêu của chúng tôi là sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) với TF-IDF và các mô hình học máy bao gồm hồi quy logistic, Naïve Bayes, SVM và phân loại bỏ phiếu dựa trên 3 mô hình trước đó. Mục tiêu của đề tài là xây dựng mô hình máy học để phân loại các bình luận thành hai loại: Positive (tích cực) hay Negative (tiêu cực). Sau đó chúng tôi cho đánh giá mô hình học máy dựa trên những thước đo: accuracy score, precision, recall và f1-score. Kết quả đạt được cho thấy các phương pháp học máy và xử lý ngôn ngữ tự nhiên có thể được sử dụng để phân loại các đánh giá phim với độ chính xác cao.**

Từ khóa—Tập dữ liệu bình luận phim ảnh IMDB, phim ảnh, học máy, phân tích cảm xúc, xử lý ngôn ngữ tự nhiên, Word stemming, TF-IDF, hồi quy logistic, Naïve Bayes, phân loại bỏ phiếu, SVM.

I. GIỚI THIỆU

Trong thời đại số hóa, phim đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của chúng ta. Đánh giá phim đã trở thành một cách để mọi người chia sẻ các quan điểm và cảm nhận của họ đối với những bộ phim. Hiểu được những cảm xúc được thể hiện trong các đánh giá phim có thể giúp các nhà làm phim, nhà quảng cáo và những người quan tâm đến ngành công nghiệp phim hiểu rõ hơn ý kiến của khán giả và đưa ra những quyết định phù hợp. Với sự phát triển lớn mạnh công nghệ, đặc biệt là mạng xã hội, đã tạo một lượng lớn dữ liệu được tạo ra từ việc đánh giá phim ảnh trực tuyến trên các nền tảng như IMDB. Những dữ liệu này chứa đựng những ý kiến và đánh giá chi tiết về các khía cạnh khác nhau của các bộ phim, bao gồm cảm xúc tích cực và tiêu cực của khán giả. Tuy nhiên, xử lý và phân tích lượng lớn dữ liệu này để thu thập được những thông tin hữu ích và những hiểu biết sâu sắc về các bộ phim là một thách thức đối với cả con người và các phương pháp truyền thống.

Bằng cách hiểu rõ hơn về cảm nhận và ý kiến của khán giả, các nhà làm phim có thể điều chỉnh nội dung, diễn xuất, âm nhạc và các yếu tố khác để tạo ra những bộ phim hấp dẫn và gợi cảm xúc mạnh. Đối với các nhà quảng cáo và đơn vị tiếp thị, phân tích cảm xúc và đánh giá phim có thể cung cấp thông tin quan trọng về mức độ phổ biến của một bộ phim và sự tương tác của khán giả với nó. Các quyết định về chiến dịch tiếp thị, việc lựa chọn đối tượng khách hàng và cách tiếp cận sẽ trở nên hiệu quả hơn khi dựa trên những thông tin chi tiết về cảm nhận và đánh giá từ người xem. Nghiên cứu này cũng mang tính ứng dụng cao, với tiềm năng để phát triển các công cụ và dịch vụ phân tích cảm xúc và đánh giá phim tự động. Các công ty trong ngành công nghiệp phim có thể tận dụng những công cụ này để thu thập phản hồi từ khán giả nhanh chóng và hiệu quả hơn, giúp định hình chiến lược tiếp thị và phát triển các sản phẩm phù hợp với nhu cầu của thị trường.

Trong bài nghiên cứu này, chúng tôi tập trung vào bài toán phân tích cảm xúc dựa trên tập dữ liệu bình luận phim ảnh trên trang IMDB, một nguồn thông tin quan trọng được công bố trên tạp chí ACL 2021. Mục tiêu của chúng tôi là sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) kết hợp với phương pháp biểu diễn từ vựng TF-IDF và các mô hình học máy như hồi quy logistic, Naïve Bayes và SVM để xây dựng mô hình máy học nhằm phân loại các bình luận thành hai loại: tích cực (positive) và tiêu cực (negative). Mục tiêu cuối cùng của đề tài là đánh giá hiệu suất của mô hình máy học dựa trên các thước đo như accuracy score, precision, recall và f1-score.

Xử lý ngôn ngữ tự nhiên mang những thách thức khó khăn vì nó yêu cầu phải phân tích các luồng dữ liệu (ở đây là các đánh giá phim) từ nhiều góc nhìn, điều này đòi hỏi rất nhiều thời gian và công sức từ con người. Nguồn dữ liệu có thể là có cấu trúc, bán cấu trúc hoặc không cấu trúc tùy thuộc vào nền tảng lưu trữ. Máy móc có thể phân tích những dữ liệu phân tán này bằng việc sử dụng những từ hay cụm từ khóa có liên quan tới nghĩa của cả câu, qua đó có thể hiểu được nghĩa của văn bản. Phân tích cảm xúc đối với ngôn ngữ tự nhiên ở dây có thể được thực hiện bằng nhiều cách, bằng các hệ thống dựa trên quy tắc, tự động hay kết hợp bằng các quy tắc thủ công và các kỹ thuật học máy. Đầu tiên trong quá trình nghiên cứu mô hình học máy này, chúng tôi muốn đề cập đến quá trình phân tích dữ liệu và xử lý để có thể dễ dàng sử dụng cho các mô hình học máy.

Việc phân tích dữ liệu thường để kiểm tra tập dữ liệu cần phải được đảm bảo rằng các nhãn được phân bố với tỉ lệ phù hợp để đưa vào huấn luyện và kiểm thử. Có thể sử dụng một số phương pháp thống kê đối với những từ được chứa trong dữ liệu, từ đó đưa ra những nhận xét khách quan để đánh giá xem tập dữ liệu có phù hợp để huấn luyện và kiểm thử hay không. Những bước này cần được thực hiện đầy đủ và cẩn thận để đảm bảo tập dữ liệu có thể phù hợp với nhu cầu sử dụng sau khi hoàn thành huấn luyện mô hình học máy. Việc xử lý dữ liệu cũng rất quan trọng vì nó đòi hỏi chúng ta phải quyết định dạng chuẩn của dữ liệu phù hợp với mô hình học máy. Chúng ta cần phải xác định loại bỏ những ký tự không liên quan đến yêu cầu cũng như những từ ngữ mang ý nghĩa mơ hồ, không ảnh hưởng đến ý nghĩa của câu trong thực tế (stop words). Tuy nhiên trong một vài trường hợp cũng cần phải giữ lại những stop words khi chúng bổ sung ý nghĩa cho câu. Ngoài ra, còn một số cách xử lý dữ liệu giúp tăng hiệu suất huấn luyện mô hình như word stemming, lemmatize, tokenize, part-of-speech, text normalize,… Bước tiếp theo là biến đổi những dữ liệu của chúng ta thành những tập hợp đặc trưng số học để chắc chắn những mô hình phân loại có thể hiểu được.

Ở đây, chúng ta có thể sử dụng những phương pháp Word Embedding như TF-IDF, bag of words (BOW), Word2Vec hay GloVe. Qua phương pháp này, chúng ta có thể biến đổi những từ ngữ sang không gian vector, biểu diễn được mối liên hệ, sự tương đồng về mặt ngữ nghĩa đối với dữ liệu chung. Ngoài ra có thể sử dụng mô hình chủ đề, ví dụ như phân bổ dirichlet tiềm ẩn (LDA) thuộc lớp mô hình sinh (generative model) xác định tập hợp các chủ đề được biểu diễn bởi tập hợp các từ.

Tiếp đến, ta sử dụng những mô hình phân loại khác nhau, có thể là những mô hình học có giám sát như hồi quy logistic, Naïve Bayes hay SVM. Ngoài ra có thể kể đến những mô hình kết hợp có thể được sử dụng đối với công việc phân loại như

phân loại bỏ phiếu hay XGBoost. Sau khi chọn được những model cũng như đã xử lý những dữ liệu sao cho phù hợp với model, chúng ta tiến hành fine-tuning để mô hình học máy đạt hiệu suất cao cho một công việc cụ thể, phù hợp với nhu cầu sử dụng. Ở bước này, ta có thể áp dụng Data Augmentation, Transfer Learning hay sử dụng Hyperparameter Tuning, tìm ra những siêu tham số trong mô hình thông qua Gridsearch hoặc Randomsearch giúp mô hình đạt hiệu suất cao nhất đối với bài toán. Sau cùng, chúng ta sẽ đánh giá lại mô hình học máy bằng những thước đo như accuracy score, recall, precision hay f1-score. Việc đánh giá mô hình học máy giúp ta đảm bảo rằng mô hình hoạt động hiệu quả và không có vấn đề nào xảy ra.

Trong bài nghiên cứu này, chúng tôi muốn trình bày nội dung như đã được cấu trúc ở trên. Phần II sẽ trình bày và phân tích sơ bộ về tập dữ liệu có sẵn, Phần III sẽ là về những phương pháp trích xuất đặc trưng trong xử lý ngôn ngữ tự nhiên được áp dụng, từ việc xử lý dữ liệu tới việc chọn mô hình phân loại và fine-tuning chúng. Phần IV sẽ là sơ lược về những mô hình học máy được sử dụng trong bài nghiên cứu, từ việc lựa chọn những mô hình phân loại đến việc tìm những tham số tối ưu cho mô hình, giúp mô hình hoạt động hiệu quả nhất. Phần V bao gồm những đánh giá chung thông qua những thước đo cơ bản đã được nêu ở trên, từ đấy rút ra những kết luận về nghiên cứu và những cải thiện cho những nghiên cứu tiếp theo trong tương lai sẽ được đề cập ở phần VI.

II. Dữ liệu

Phần nội dung này sẽ cung cấp sơ bộ về bộ dữ liệu IMDB Movie Reviews được cung cấp trên tạp chí ACL 2011 với tên gọi “IMDB Movie Reviews Dataset”.

A. *Mô tả sơ bộ về bộ dữ liệu*

Bộ dữ liệu "IMDB Movie Reviews Dataset" là một tập dữ liệu quan trọng và nổi tiếng, được sử dụng phổ biến trong những bước đầu tiếp cận với việc phân tích cảm xúc của ngôn ngữ tự nhiên. Bộ dữ liệu này bao gồm tổng cộng 50000 đánh giá phim được thu thập từ trang web IMDB. Mỗi đánh giá có hai nhãn: tích cực (positive) hoặc tiêu cực (negative), với số lượng nhãn tích cực và tiêu cực được cân bằng (mỗi nhãn bao gồm 25000 đánh giá. Điều này làm cho bộ dữ liệu trở thành một tài nguyên quý giá để huấn luyện và đánh giá các mô hình phân loại cảm xúc của ngôn ngữ tự nhiên.

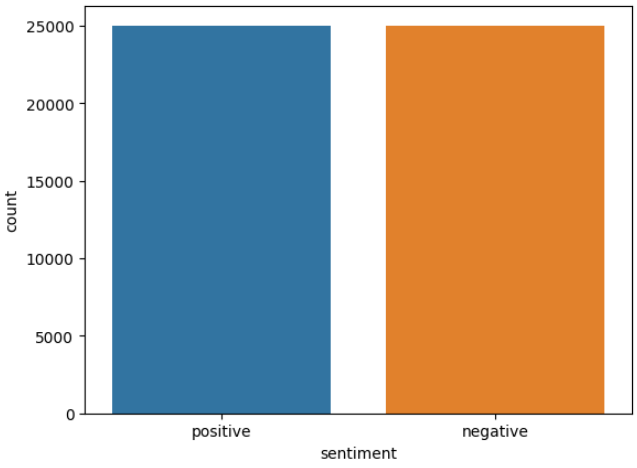

Fig. 1. Biểu đồ thống kê về số lượng nhãn tích cực và tiêu cực trong tập dữ liệu, IMDB Movie Reviews Dataset.

B. *Mục đích và ý nghĩa của bộ dữ liệu*

Bộ dữ liệu IMDB Movie Reviews có vai trò quan trọng trong việc nghiên cứu cũng như phát triển các phương pháp phân tích cảm xúc trong ngôn ngữ tự nhiên. Việc hiểu và phân loại cảm xúc trong đánh giá phim không chỉ cung cấp những thông tin khác nhau mà còn có thể áp dụng rộng rãi trong nhiều lĩnh vực khác nhau như đánh giá sản phẩm, phân tích tư duy của người dùng, hoặc xây dựng hệ thống gợi ý phim (recommend system). Bộ dữ liệu IMDB Movie Reviews cung cấp một cơ sở đáng tin cậy để xây dựng và đánh giá các mô hình phân tích cảm xúc, cụ thể ở đây là bài nghiên cứu của chúng tôi.

C. *Phân tích sơ bộ về bộ dữ liệu*

Bộ dữ liệu IMDB Movie Reviews chứa các đánh giá phim với tổng 2470 từ là số lượng từ nhiều nhất và 4 từ là số lượng từ thấp nhất cho mỗi đánh giá phim. Sau khi thống kê bằng histogram, chúng tôi nhận thấy đối với các bình luận nhiều chữ đa số thuộc về những nhãn cảm xúc tích cực (positive), còn đối với những bình luận ít chữ thì thuộc về những nhãn tiêu cực (negative).

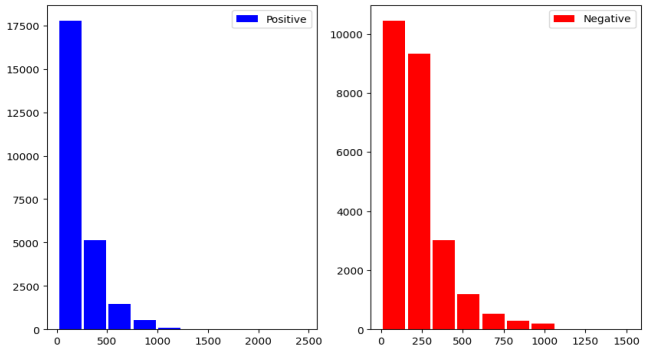

Fig. 2. Histogram biểu diễn số từ trong đánh giá phim được gán nhãn cảm xúc tương ứng.

Sau quá trình phân tích, chúng tôi nhận thấy các đánh giá của dữ liệu có chứa nhiều ký tự vô nghĩa hoặc có khả năng dẫn tới việc phân loại sai sau khi được đưa vào mô hình, nhóm chúng tôi tiến hành tiền xử lý dữ liệu. Bước tiếp theo, mục tiêu của chúng tôi đề ra là phải hoàn thành loại bỏ những ký tự vô nghĩa đó và những stop words không quan trọng trong phân tích ngôn ngữ tự nhiên, từ đó góp phần tăng hiệu suất hoạt động của những mô hình phân loại

D. *Tiền xử lý dữ liệu*

Như đã được nêu ở trên, phần dữ liệu về những đánh giá phim ở trong bộ dữ liệu ban đầu vẫn chưa sẵn sàng để được đưa vào mô hình học máy để tiến hành huấn luyện phân loại được. Giải thích của việc này phần lớn là vì mô hình phân loại học máy cần dữ liệu ở dạng đã được quy định từ trước chứ không phải ở dạng đoạn văn ban đầu. Hơn nữa, do phần lớn dữ liệu ở đây được thu thập ở trên mạng internet nên có thể có những tag HTML hay những từ viết tắt. Những kỹ thuật tiền xử lý dữ liệu được chúng tôi sử dụng trong bài nghiên cứu bao gồm:

- Loại bỏ những từ viết hoa sử dụng hàm lower() - Việc này giúp làm giảm số lượng từ khác nhau trong bộ dữ liệu, qua đó góp phần làm giảm số chiều của dữ liệu và cải thiện hiệu suất hoạt động của mô hình học máy.
- Loại bỏ những tag HTML - Như đã được nêu ở trên, dữ liệu có nguồn gốc là từ mạng internet, nên có thể tồn tại những tag HTML cần phải bị loại bỏ vì chúng

không mang ý nghĩa để phân loại cảm xúc, nếu những tag HTML này được giữ lại sẽ trở thành những vector gây hại cho lưu trữ bộ nhớ cũng như việc phân loại cảm xúc có thể bị sai lệch.

- Loại bỏ những URL (web links) - Trong những đánh giá có thể tồn tại những đường dẫn không liên quan, không ảnh hưởng tới nội dung của đánh giá. Việc loại bỏ những đường dẫn này cũng giúp làm giảm số chiều của dữ liệu, góp phần cải thiện hiệu suất hoạt động của mô hình học máy.
- Loại bỏ những hastag (@ hoặc #) - Trong những đánh giá có tồn tại những hastag này, chúng không liên quan đến nội dung của đánh giá. Tương tự, việc loại bỏ những hastag này cũng sẽ giúp làm giảm số chiều của dữ liệu, góp phần cải thiện hiệu suất hoạt động của mô hình học máy.
- Phân tích từ vựng - Chúng tôi sử dụng hàm phân tách văn bản đầu vào thành một danh sách các từ sử dụng word_tokenizer() từ thư viện nltk chuyên dùng để xử lý ngôn ngữ tự nhiên. Điều này được thực hiện để chia nhỏ văn bản đầu vào thành các từ riêng lẻ, giúp cho việc phân tích văn bản được thực hiện bởi các mô hình diễn ra dễ dàng hơn, góp phần gia tăng hiệu suất hoạt động của mô hình.
- Loại bỏ những từ dừng (stop words) - Thông qua việc sử dụng một danh sách các stop words có từ trước, chúng tôi tiến hành loại bỏ những từ như là “is”, “am”, are”, “the”, và những từ tương tự bởi chúng không mang lại ý nghĩa cho đánh giá phim. Những từ này chỉ được sử dụng để hỗ trợ các từ mang ý nghĩa chính. Loại bỏ những từ này ngoài giúp giảm chiều dữ liệu cũng một phần gia tăng hiệu suất hoạt động của mô hình học máy.
- Loại bỏ những đánh giá trùng lặp - Trong bộ dữ liệu IMDB Movie Reviews, có tồn tại những đánh giá trùng lặp với nhãn giống nhau. Trong quá trình tiến hành phân tích bộ dữ liệu, chúng tôi phát hiện có 421 đánh giá bị trùng lặp. Những đánh giá này phần lớn có số lượng từ ít và mang những nhãn cảm xúc giống nhau, loại bỏ những đánh giá sẽ góp phần làm giảm kích thước bộ dữ liệu.
- Word Stemming (rút gọn từ) - Đây là một phương pháp được sử dụng trong xử lý ngôn ngữ tự nhiên nhằm đưa các từ về dạng gốc (stem) nhằm giảm độ phức tạp và đồng nhất hóa từ vựng. Ở trong bài nghiên cứu này, chúng tôi sử dụng Porter Stemmer, đây là một thuật toán rút gọn từ bằng cách loại bỏ các hậu tố từ được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên. Áp dụng kỹ thuật này giúp chúng tôi giảm chiều dữ liệu, đưa chúng về những từ khóa nhất quán hơn, chỉ quan tâm tới nghĩa của các từ, giúp tránh được các trường hợp riêng biệt của từ. Kỹ thuật này giúp giảm chiều dữ liệu từ đó cải thiện tốc độ học cũng như hiệu suất của mô hình.

Sau khi tiến hành những kỹ thuật tiền xử lý dữ liệu ở trên, chúng tôi phân tích số lượng những từ được lặp lại thường xuyên trong những đánh giá được gán nhãn tích cực hoặc tiêu cực, biểu diễn top 15 những từ đối với mỗi nhãn trong biểu đồ. Kết quả cho thấy hầu hết những từ xuất hiện phổ biến nhất (top 5) của cả 2 nhãn giống với nhau, điều này có thể xảy ra khi phần lớn liệu mang ý nghĩa trung lập (neutral), những từ mang ý nghĩa tổng quát, không diễn tả rõ ý nghĩa tiêu cực hay tích cực đối với bộ phim. Ngoài ra cũng có thể do tập dữ liệu hiện tại vẫn thiếu sự đa dạng trong vốn từ, dẫn tới thiếu những từ ngữ mang tính cụ thể, diễn đạt mạnh mẽ đến cảm xúc tích cực hay tiêu cực của nhãn cảm xúc. Để cải thiện bộ dữ liệu, ta có thể cho thêm nhãn trung lập (neutral), mở rộng tập dữ liệu rộng hơn, giúp đa dạng hơn về vốn từ có đủ khả năng diễn đạt rõ ý nghĩa của nhãn cảm xúc.

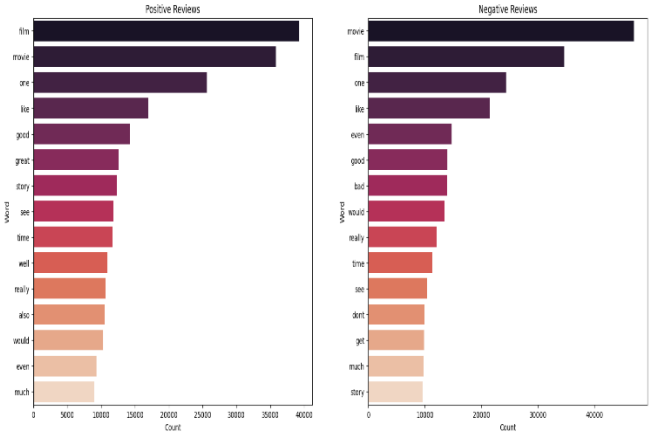

Fig. 3. Biểu đồ thống kê top 15 từ xuất hiện nhiều nhất trong đánh giá phim tương ứng với từng nhãn cảm xúc.

E. Phân phối các nhãn và phân chia bộ dữ liệu thành các tập nhỏ hơn

Trước khi bắt đầu xử lý dữ liệu từ bộ dữ liệu IMDB Movie Reviews, có một bước quan trọng chính là phân phối nhãn và phân chia dữ liệu. Với bộ dữ liệu này, nhãn cảm xúc của những bình luận được gán thành positive (tích cực) hoặc negative (tiêu cực). Đồng thời, việc sở hữu sự cân bằng giữa các nhãn (50% positive và 50% negative) đóng góp vai trò quan trọng để đảm bảo mô hình học máy được huấn luyện trên một tập dữ liệu có tính ổn định ở phân bố của các nhãn, góp phần loại bỏ sự thiên vị giữa các nhãn khi đưa vào quá trình huấn luyện, đây là một bộ dữ liệu có tính công bằng cao.

Sau khi phân phối nhãn dữ liệu, tiếp theo là phân chia dữ liệu thành các tập huấn luyện và tập kiểm thử. Phân chia này giúp đánh giá hiệu suất và độ chính xác của mô hình. Thông thường, một tỷ lệ phân chia phổ biến là 80-20 hoặc 70-30, tức là 80% (70%) dữ liệu cho tập huấn luyện và 20% (30%) dữ liệu cho tập kiểm thử. Tỷ lệ này phụ thuộc vào kích thước của tập dữ liệu và đặc điểm của bộ dữ liệu. Trong bài nghiên cứu này, sau khi áp dụng kiểm tra ngẫu nhiên 5 lần random_state với model đại diện là hồi quy Logistic và mỗi tỷ lệ kiểm thử đã được nêu ở trên, kết quả cho thấy với test_size = 0.3 thì hiệu suất huấn luyện cho ra là cao nhất. Qua đấy, chúng tôi tiến hành sử dụng tỷ lệ huấn luyện - kiểm thử là 70-30.

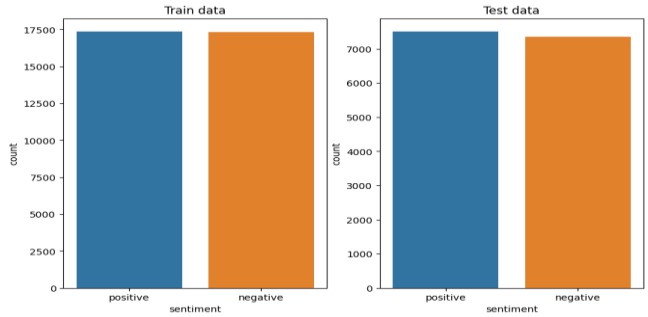

Fig. 4. Thống kê về số lượng các nhãn cảm xúc trong tập huấn luyện và tập kiểm thử đã được chia theo tỷ lệ 70-30.

III. TRÍCH XUẤT ĐẶC TRƯNG

Trong xử lý ngôn ngữ tự nhiên, trích xuất đặc trưng là quá trình chuyển đổi văn bản thành một tập hợp các đặc trưng số học để có thể được sử dụng để huấn luyện mô hình học máy. Các phương pháp trích xuất đặc trưng có thể được áp dụng ở đây đã được kể đến ở phần giới thiệu như TF-IDF, bag of words, word2vec hay GloVe. Trong bài nghiên cứu này, nhóm chúng tôi sử dụng TF-IDF như phương pháp trích xuất đặc trưng dữ liệu đối với những đánh giá phim trong bộ dữ liệu IMDB Movie Reviews.

Như nội dung được trình bày tại chương về trích xuất đặc trưng trong cuốn sách [2], TF-IDF (Term frequency-Inverse document frequency) là một phương pháp đánh trọng số các từ xuất hiện trong văn bản. Phương pháp này thường sử dụng để biểu diễn các văn bản dưới dạng các vector (cho mục đích phân loại, phân cụm, trực quan hóa, truy xuất,…). Giả sử: $T = \{t_1,...,t_n\}$ là tập hợp của tất cả các từ xuất hiện trong tập dữ liệu đang xét. Khi đó, một văn bản d_i được biểu diễn bởi vecto có n chiều giá trị thực $x_i = (x_{i1},...,x_{in})$ với mỗi đối tượng tương ứng với một từ có thể có trong T.

Phương pháp này cân nhắc tới 2 yếu tố chính là tần số xuất hiện của từ trong văn bản và tần suất nghịch đảo của từ trong tập dữ liệu:

- Tần số xuất hiện của từ (TF): đo lường tần số suất hiện của từ trong văn bản, tần số càng cao, từ đó xuất hiện càng nhiều trong văn bản.
- Tần suất nghịch đảo của từ (IDF): đo lường mức độ quan trọng của từ trong toàn bộ tập dữ liệu. Tần suất nghịch đảo càng cao có nghĩa rằng từ đó xuất hiện ít trong toàn bộ tập dữ liệu.

Trọng số của x_{ij} tương ứng với từ t_j trong văn bản d_i thường là 1 tích của ba giá trị. Một phần phụ thuộc vào tần số xuất hiện của từ t_j trong văn bản d_i (IDF), một phần phụ thuộc vào mức độ quan trọng của t_j trong toàn bộ tập dữ liệu. Thông thường, trọng số TF-IDF được tính bằng công thức:

$$\text{TF-IDF}(t_j, d_i) = \text{TF}_{t_j,d_i} \times \text{IDF}(t_j) \quad (1)$$

Với $TF_{tj,di}$ là tần số xuất hiện của từ t_j trong d_j, được tính bằng nhiều cách khác nhau, ví dụ như đếm số lần xuất hiện của từ t_j trong d_i hoặc sử dụng các phương pháp chuẩn hóa tần số.

$IDF(t_j)$ là tần suất nghịch đảo của từ t_j. Nó được tính bằng cách lấy tỷ lệ giữa tổng số tài liệu trong tập dữ liệu (N) và số từ chứa từ t_j ($DF(t_j)$), sau đó lấy log cơ số e của kết quả. Công thức chính xác là:

$$\text{IDF}(t_j) = \log_e\left(\frac{N}{\text{DF}(t_j)}\right) \quad (2)$$

Sau khi tiến hành quá trình trích xuất đặc trưng bằng việc sử dụng TfidfVectorizer của thư viện sklearn, kết quả cho ra là ma trận có kích thước (49578, 221768) biểu thị cho 49578 đánh giá phim được lọc ra trong tập dữ liệu với 221768 đặc trưng được trích xuất từ các đánh giá phim trong tập dữ liệu.

IV. MÔ HÌNH HỌC MÁY

Trong quá trình nghiên cứu, chúng tôi đã sử dụng hồi quy Logistic, Support Vector Machine, phân loại Naïve bayes. Từ những mô hình học máy ở trên, chúng tôi sử dụng phân loại bỏ phiếu, sử dụng cả hard_vote lẫn soft_vote để có những cách nhìn khách quan về cách mô hình phân loại nhãn cảm xúc. Ngoài ra, chúng tôi sử dụng những mô hình này để dự đoán nhãn đầu ra ở mảng nhãn cảm xúc lẫn mảng chứa tỉ lệ xác suất dự đoán nhãn cảm xúc, từ đấy đưa ra những đặc trưng trong cách dự đoán của các mô hình học máy.

A. Hồi quy Logistic

Hồi quy logistic [3] là một phương pháp thông kê được sử dụng để biểu thị mối quan hệ giữa biến kết quả nhị phân, một hoặc nhiều biến dự đoán. Ý tưởng cơ bản của hồi quy Logistic là sử dụng hàm Logistic để ánh xạ các biến dự báo vào một thang đo có xác suất từ 1 đến 0, từ đó có thể dự báo xác suất biến kết quả có thể nhận được giá trị mong muốn.

Hàm logistic được xác định bởi phương trình:

$$p = \frac{1}{1+e^{-z}} \quad (3)$$

Trong đó p là xác suất biến kết quả nhận giá trị quan tâm, z là tổ hợp tuyến tính của các biến dự báo. Các hệ số của mô hình hồi quy logistic được ước lượng bằng phương pháp ước lượng hợp lý cực đại (Maximum likelihood estimation). Trong đó, đánh giá xem các giá trị hệ số nào sẽ cung cấp khả năng làm cho xác suất dự báo của biến kết quả gần nhất có thể với các giá trị thực tế của biến kết quả trong tập dữ liệu. Hiệu suất của mô hình hồi quy logistic có thể được đánh giá bằng các chỉ số đo lường sự phù hợp, như là độ lệch, AIC hoặc BIC.

Trong bài nghiên cứu, nhóm chúng tôi sử dụng mô hình học máy hồi quy Logistic của thư viện sklearn [4]. Sau quá trình tinh chỉnh sử dụng GridsearchCV để tinh chỉnh mô hình với tham số C có giá trị [0.001, 0.01, 0.1, 1, 10] thì kết quả cho thấy tham số C với giá trị 10. Tham số này đóng vai trò thay đổi sự chặt chẽ của việc phân loại trong mô hình hồi quy. Giá trị C càng nhỏ thì mô hình có xu hướng phân loại linh hoạt hơn, cho phép các điểm dữ liệu bị phân loại sai giữa các lớp. Giá trị C càng lớn thì mô hình có xu hướng phân loại chặt chẽ hơn, mô hình sẽ ưu tiên tìm một ranh giới quyết định rõ ràng hơn giữa các lớp.

B. Support Vector Machine

Support Vector Machine là một thuật toán máy học được sử dụng rộng rãi trong các tác vụ phân loại và dự đoán. Ý tưởng cơ bản của SVM là tìm một ranh giới quyết định tốt nhất để phân chia các điểm dữ liệu của các lớp khác nhau trong không gian đặc trưng.

Trên thực tế, thuật toán SVM tìm ra một mặt phẳng hoặc siêu phẳng trong không gian đặc trưng, tối đa hóa khoảng cách giữa các điểm dữ liệu gần nhất của các lớp khác nhau tới mặt phẳng đó Những điểm dữ liệu gần nhất được gọi là các vectơ hỗ trợ (support vectors), tên gọi của thuật toán SVM cũng được gọi theo từ đó.

Để tìm ra mặt phẳng chia tốt nhất, SVM phải tối thiểu hóa một hàm mất mát. Trong trường hợp của SVM, hàm mất mát thường kết hợp giữa việc tối thiểu hóa lề (margin) và giảm thiểu sự phân loại sai số (classification error). Tham số quan trọng trong SVM là tham số điều chỉnh độ quan trọng giữa hai mục tiêu này, gọi là tham số C (giống với tham số được điều chỉnh ở mô hình hồi quy Logistic đã được nêu ở trên). Tham số C càng lớn, mô hình SVM sẽ cố gắng giảm thiểu sự phân loại sai số hơn, dễ dẫn đến overfitting. Ngược lại thì mô hình SVM sẽ tập trung tạo ra một ranh giới lề lớn hơn, có thể dẫn đến underfitting

Để ước lượng tham số C tốt nhất, một phương pháp phổ biến là sử dụng kỹ thuật cross-validation và tìm kiếm thông qua các giá trị C khác nhau để đánh giá hiệu suất của mô hình SVM trên tập dữ liệu kiểm tra. GridSearchCV là một phương thức phổ biến để tìm kiếm siêu tham số tốt nhất trong linearSVC của thư viện sklearn [5], tương tự như hồi quy Logistic đã được đề cập ở trên, mô hình được tinh chỉnh với tham số C[0.1, 1, 10, 100] và loss['hinge', 'squared_hinge']. Kết quả thu được là mô hình hoạt động tốt nhất với tham số C với giá trị 1 và cách tính hàm mất mát (loss) là hinge.

C. *Naïve Bayes*

Naïve Bayes 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 dưa trên xác suất để dự đoán hay phân lớp nhãn của một mẫu dựa trên xác suất tiên nghiệm và xác suất hậu nghiệm. Naïve Bayes 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:

$$P(y|x) = \frac{P(x|y)\cdot P(y)}{P(x)} \quad (4)$$

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.

Trong bài nghiên cứu này, nhóm chúng tôi sử dụng thuật toán multinomialNB của thư viện sklearn [6]. Trong multinomial Naïve Bayes (multimonialNB), ta giả định các đặc trưng là biến rời rạc và tuân theo phân phối đa thức (multinomial). Trong mô hình multinomialNB thuộc thư viện sklearn, tham số alpha là một tham số quan trọng, tham gia điều chỉnh mức độ ảnh hưởng của xác suất tiên nghiệm của các đặc trưng khi biết lớp trong quá trình tính toán. Nếu giá trị của alpha lớn, mô hình sẽ đặt mức độ ảnh hưởng cao cho xác suất tiên nghiệm và giảm khả năng bị overfitting. Ngược lại, với alpha nhỏ thì mô hình sẽ tập trung vào dữ liệu huấn luyện và có khả năng bị overfitting.

Trong quá trình nghiên cứu, qua sử dụng GridSearchCV để điều chỉnh tham số alpha[0.001, 0.01, 0.1, 1, 10]. Kết quả thu được là với tham số 1, mô hình multinomialNB hoạt động tốt nhất.

D. *Phân loại bỏ phiếu (Voting classifier)*

Phân loại bỏ phiếu là một kỹ thuật phân loại được sử dụng để kết hợp các dự đoán từ nhiều mô hình khác nhau. Kỹ thuật này có hai phương pháp phân loại bỏ phiếu chính: hard vote (bỏ phiếu cứng) và soft vote (bỏ phiếu mềm).

Trong phân loại bỏ phiếu cứng (hard vote), nhiều mô hình được huấn luyện trên các tập dữ liệu con khác nhau hoặc với các siêu tham số khác nhau. Khi phân loại một mẫu mới, mỗi mô hình đưa ra một dự đoán riêng và nhãn cuối cùng được quyết định bằng cách bỏ phiếu. Nhãn cuối cùng được chọn là nhãn có số phiếu bỏ nhiều nhất từ các mô hình.

Trong phân loại bỏ phiếu mềm (soft vote), mỗi mô hình tính toán xác suất dự đoán cho mỗi lớp cho mẫu mới. Nhãn cuối cùng được quyết định bằng cách tính trung bình hoặc trung vị của các xác suất dự đoán từ các mô hình. Nhãn có xác suất cao nhất sẽ được chọn là nhãn cuối cùng.

Trong bài nghiên cứu này, chúng tôi sử dụng cả bỏ phiếu cứng và bỏ phiếu mềm (với mô hình VotingClassifier thuộc thư viện sklearn [7]) với các mô hình thành phần là ba mô hình đã được giới thiệu trước đó, hồi quy Logistic, SVM và Naïve Bayes.

V. KẾT QUẢ NGHIÊN CỨU

Sau khi thực hiện những phân loại với những mô hình học máy được nêu ở trên, chúng tôi kiểm tra hiệu năng của mô hình sử dụng những thước đo như accuracy score, recall, precision và f1-score. Kết quả nghiên cứu sẽ được tóm tắt trong bảng sau.

TABLE I. BẢNG ĐO HIỆU SUẤT NHỮNG THUẬT TOÁN HỌC MÁY TRONG BÀI NGHIÊN CỨU

Thuật toán máy học	**Thước đo**			
	Accuracy	***Precision***	***Recall***	***F1-Score***
Logistic Regression	0.8942	0.8944	0.8941	0.8942
SVM	0.8914	0.8918	0.8913	0.8914
Naïve Bayes	0.8644	0.8645	0.8645	0.8644
Voting clf (hard)	0.8946	0.8948	0.8945	0.8945
Voting clf (soft)	0.8961	0.8961	0.8960	0.8960

Qua bảng đo hiệu suất ở trên, ta có thể rút ra những nhận xét về từng mô hình như sau:

- Logistic Regression: Thuật toán Logistic Regression đạt được mức độ chính xác (accuracy) cao nhất trong số các thuật toán được thử nghiệm, đạt 89.42%. Các thước đo precision, recall và F1-Score của nó đều gần như bằng nhau và cao.
- SVM: Thuật toán SVM đạt mức độ chính xác 89.14%, gần tương đương với Logistic Regression. Precision, recall và F1-Score của SVM cũng gần như bằng nhau và cao.
- Naïve Bayes: Thuật toán Naïve Bayes đạt mức độ chính xác 86.44%. Precision, recall và F1-Score của Naïve Bayes cũng gần như bằng nhau, nhưng thấp hơn so với Logistic Regression và SVM.
- Voting clf (hard): Khi sử dụng phương pháp phân loại bỏ phiếu cứng (hard vote), mô hình kết hợp các thuật toán máy học đạt mức độ chính xác 89.46%, gần tương đương với Logistic Regression và SVM. Các thước đo precision, recall và F1-Score cũng tương tự nhau và cao.
- Voting clf (soft): Khi sử dụng phương pháp phân loại bỏ phiếu mềm (soft vote), mô hình kết hợp các thuật toán máy học đạt mức độ chính xác cao nhất là 89.61%. Precision, recall và F1-Score đều gần như bằng nhau và cao.

Tổng quan, các thuật toán máy học đều đạt được mức độ chính xác khá cao và có độ cân bằng tương đối giữa các thước đo precision, recall và F1-Score.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong nghiên cứu này, chúng tôi đã tập trung vào bài toán phân tích cảm xúc dựa trên tập dữ liệu bình luận phim ảnh từ trang IMDB. Mục tiêu của chúng tôi là sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) kết hợp với các mô hình học máy như hồi quy logistic, Naïve Bayes, SVM và phân loại bỏ phiếu để xây dựng một mô hình phân loại các bình luận thành hai nhãn: tích cực (positive) hoặc tiêu cực (negative).

Chúng tôi đã sử dụng phương pháp TF-IDF để trích xuất đặc trưng từ văn bản và áp dụng các mô hình học máy để huấn luyện và đánh giá mô hình. Các thước đo như accuracy score, precision, recall và F1-score đã được sử dụng để đánh giá hiệu suất của các mô hình.

Kết quả cho thấy cả hai phương pháp học máy và xử lý ngôn ngữ tự nhiên đều cho kết quả phân loại chính xác và có hiệu suất cao trong việc phân loại đánh giá phim. Mô hình học máy như hồi quy logistic, Naïve Bayes và SVM đạt được độ chính xác tương đối cao, trong khi phân loại bỏ phiếu dựa trên các mô hình trước đó cho kết quả tốt hơn.

Trong quá trình nghiên cứu, chúng tôi đã có tạo một web demo để kiểm tra và đánh giá hiệu suất những model trên thông qua việc gán nhãn cho những đánh giá phim tự nhập bằng tay bằng streamlit. Thông qua web demo này, chúng tôi mong có thể đánh giá tốt hơn những lỗi có thể gặp trong việc gán nhãn của những mô hình học máy này.

Trong tương lai, có một số hướng phát triển tiềm năng cho nghiên cứu này. Đầu tiên, có thể tăng cường các kỹ thuật xử lý ngôn ngữ tự nhiên bằng cách sử dụng các phương pháp khác nhau như word embedding, mô hình ngôn ngữ tiên đoán (pre-trained language models) để cải thiện hiệu suất phân loại. Thứ hai, có thể thử nghiệm với các mô hình học máy khác nhau hoặc sử dụng mô hình học sâu (deep learning) để đạt được kết quả tốt hơn. Ngoài ra có thể thử nghiệm phân loại các lớp với nhiều nhãn hơn (có thể thêm với nhãn trung tính neutral).

[1] A. I. R. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, *Learning Word Vectors for Sentiment Analysis*. 2011, pp. 142–150.

[2] W. T. B. Uther *et al.*, "TF–IDF," in *Springer eBooks*, 2011, pp. 986–987. doi: 10.1007/978-0-387-30164-8_832.

[3] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, Jan. 2011, doi: 10.1002/widm.8.

[4] "sklearn.linear_model.LogisticRegression".Scikit-learn:Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5] "sklearn.svm.LinearSVC".Scikit-learn:Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] "sklearn.naive_bayes.MultinomialNB".Scikit-learn:Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[7] "sklearn.ensemble.VotingClassifier". Scikit-learn:Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.