BÀI TOÁN PHÂN TÍCH CẢM XÚC CỦA BÌNH LUẬN VỀ PHIM DỰA TRÊN SPARK

Hồ Anh Dũng[1,*], Nguyễn Hữu Trường[1,*], Võ Đình Tứ[1,*], and Đỗ Trọng Hợp[1,†]

[1]Trường Đại Học Công Nghệ Thông Tin
Đại Học Quốc Gia Thành Phố Hồ Chí Minh
[*]*{18520630, 18521564, 18521589}@gm.uit.edu.vn*
[†]*{hopdt}@uit.edu.vn*

Tóm tắt nội dung Phân tích tự động trong các đánh giá của khách hàng trực tuyến đang là một trong những chủ đề nghiên cứu nóng gần đây. Khi số lượng bình luận đánh giá ngày càng nhiều, điều cần thiết là phải phát triển một mô hình phân tích đánh giá hiệu quả có khả năng đánh giá hiệu suất hoạt động của sản phẩm dựa trên các bình luận bằng cách xác định cảm xúc cho mỗi bình luận đó là tiêu cực hay tích cực. Nếu làm tốt được điều này, lượng thông tin thu được sẽ góp phần cải thiện hiệu suất kinh doanh một cách đáng kể. Một bài toán cơ bản và là nền tảng giải quyết vấn đề trên chính là bài toán phân tích cảm xúc dựa trên các review đã được nghiên cứu rất nhiều ở thời điểm hiệm tại (Sentiment analysis). Bài toán này tập trung vào việc nhận dạng các bình luận và từ đó đưa ra đánh giá xu hướng cảm xúc đối với bình luận đó. Trong bài báo này chúng tôi trình bày và giải quyết bài toán phân loại cảm xúc của bình luận về phim ảnh dựa trên một framework là Spark, sử dụng dữ liệu lớn (BigData). Bộ dữ liệu chúng tôi sử dụng trong bài báo này có tên là IMDb Largest Review Dataset. Kết quả thực nghiệm cho thấy rằng mô hình Logistic Regression cho kết quả tốt nhất cho bài toán với 79% cho nhiệm vụ phân loại cảm xúc của các bình luận dựa trên Spark.

Keywords: Phân tích cảm xúc các bình luận phim · Phân loại cảm xúc dựa trên Spark · Spark · Bigdata.

1 Giới thiệu

Ngày nay, với sự phát triển nhanh chóng của các nội dung do người dùng tạo ra trên internet, phân tích cảm xúc tự động trong các đánh giá của khách hàng trực tuyến đã trở thành một chủ đề nghiên cứu nóng gần đây. Với số lượng đa dạng các loại sản phẩm và dịch vụ được đánh giá trên nhiều trang web, điều cần thiết là phải phát triển một mô hình phân tích cảm xúc hiệu quả có khả năng trích xuất các khía cạnh của sản phẩm mà người dùng đề cập và xác định cảm xúc mà khách hàng đánh giá về sản phẩm đó. Từ đó có thể khai thác thông tin này để lên kế hoạch kinh doanh phù hợp hơn, cải thiện chất lượng sản phẩm cũng như dịch vụ đi kèm để phù hợp hơn với người tiêu dùng. Đó là lý do vì sao bài toán phân tích cảm xúc là một chủ đề được quan tâm rất nhiều. Ngoài ra bài toán còn được phát triển trên những nền tảng framework phát triểu khác nhau, ở đây chúng sử dụng Spark[1] trong phân tích dữ liệu lớn.

[1] https://spark.apache.org/

Bài toán của chúng tôi thực hiện lần lượt hai nhiệm vụ chính: phân loại cảm xúc của bình luận phim dựa trên Spark[2] sử dụng bộ dữ liệu lớn có sẵn, nhiệm vụ còn lại là phân tích cảm xúc của bình luận phim trên thời gian thực (Realtime). Các nội dung cụ thể của hai nhiệm vụ này sẽ được chúng tôi nêu rõ ở phần 3.2.

Để giải quyết được nhiệm vụ đầu tiên, chúng tôi tiến hành thực nghiệm trên bộ dữ liệu lớn có tên là IMDb Largest Review Dataset. Chúng tôi đã áp dụng các mô hình máy học truyền thống trong Spark (Naive Bayes, Logistic Regression), tương tự một mô hình học sâu (ClassifierDLApproach) để so sánh hiệu suất trên bộ dữ liệu.

Phần còn lại của bài báo sẽ được tổ chức như sau: Phần 2 sẽ trình bày các công trình cụ thể liên quan đến bài toán của chúng tôi. Bộ dữ liệu được thực nghiệm và cách tiếp cận bài toán sẽ được nêu rõ trong phần 3. Phần 4 sẽ là quá trình chuẩn bị dữ liệu, đề xuất các kịch bản thực nghiệm và đưa ra các kết quả mô hình. Cuối cùng, phần 5 sẽ đưa ra những kết luận, khó khăn và những công việc giải quyết trong tương lai.

2 Công trình liên quan

Trong suốt thập kỷ qua, phân tích cảm xúc dựa theo bình luận sản phẩm được xem là một trong những phần quan trọng nhất trong thương mại điện tử vì nó có ứng dụng tiềm năng để phân tích phản hồi của người dùng trực tuyến dựa trên các quan điểm và nhận xét của họ [1], [2].

Vào năm 2004, M Hu và B đã đề xuất vấn đề tổng hợp ý kiến người dùng về các tính năng của sản phẩm được bán trực tuyến [3]. Vào năm 2009, [4] đề xuất hệ khuyến nghị dựa trên hồi quy, sử dụng các bình luận đánh giá của người dùng. [5] dựa trên cấu trúc ngữ pháp của mệnh đề, Tun Thura Thet và các cộng sự đã đề xuất một phương pháp tiếp cận ngôn ngữ để tính toán điểm cảm xúc cho một mệnh đề đối với các khía cạnh và bình luận đánh giá khác nhau của một bộ phim, năm 2010, Tương tự, như vậy ở năm kế tiếp (2011) [6] đã đề xuất mô hình SLDA và ASUM cho việc trích xuất các các cặp thể loại (category), tình cảm trong các bình luận về thiết bị điện tử, nhà hàng và kể cả phim ảnh. Đây chính nguồn gốc ra đời những bài toán phân tích cảm xúc khác, ở đây chúng tôi đề cập đến là bài toán Real-time Sentiment Analysis (SA).

Mặc dù có nhiều công trình về mô hình dựa trên chủ đề cho hệ thống process systems cho bài toán SA, nhưng có rất ít công trình trong các tài liệu về mô hình dựa trên các chủ đề nổi biệt hiện nay như phim ảnh, nhà hàng, thương mại điện tử,...để phân tích cảm xúc thời gian thực (realtime) trên dữ liệu truyền trực tuyến (streaming data). Nhiệm vụ phân tích cảm xúc theo thời gian thực là bắt buộc để đáp ứng các hạn chế nghiêm ngặt về thời gian và không gian để xử lý dữ liệu truyền trực tuyến một cách hiệu quả [7]. Wang và cộng sự. trong bài báo [7] đã phát triển một hệ thống phân tích tình cảm Twitter theo thời gian thực về "Chu kỳ bầu cử Tổng thống 2012" (Presidential Election Cycle) bằng cách sử dụng dữ liệu Twitter firehose với mô hình thống kê cảm xúc (statistical sentiment) và bộ phân loại Naive Bayes trên các tính năng unigram. Một bộ phân tích đầy đủ đã được phát triển để theo dõi sự thay đổi trong cảm xúc bằng cách sử dụng các quy tắc và từ khóa được chuyên gia tuyển chọn để có được bức tranh chính xác về bối cảnh chính trị trực tuyến trong thời gian thực. Tuy nhiên, những tác phẩm này trong các tài liệu hiện có thiếu sự phức tạp của các quá trình phân tích tình cảm. Mô hình phân tích tình cảm cho hệ thống của họ chỉ dựa trên các tổng hợp đơn giản để tóm tắt thống kê với những kỹ thuật tiền xử lý ngôn ngữ truyền thống đạt hiệu suất tối thiểu.

[2] https://spark.apache.org/

Các nghiên cứu gần đây hơn [8] [9] đã đề xuất các kiến trúc xử lí streaming data dựa trên dữ liệu lớn (Bigdata). Công trình đầu tiên vào năm 2015 [8] đề xuất phương pháp tiếp cận dựa trên multi-layered storm để áp dụng phân tích cảm tính trên các dữ liệu streams lớn trong thời gian thực và công trình thứ hai vào năm 2016 [9] đề xuất khung phân tích dữ liệu lớn để phân tích cảm xúc người tiêu dùng được embedded trong hàng trăm triệu bài đánh giá sản phẩm trực tuyến cũng trên thời thực. Ở công trình này, chúng tôi tiếp cận giống với đề xuất của công trình 2016, nhưng xây dựng bài toán dựa trên bộ dữ liệu phát triển của chúng tôi và thực hiện bài toán dựa trên framework Spark trong dữ liệu lớn. Cụ thể hơn, về khung phân tích dữ liệu cho bài toán dựa trên những bình luận đánh giá về phim ảnh trên trang điện tử IMDB.

3 Bộ dữ liệu và cách tiếp cận

3.1 Bộ dữ liệu

1. **Thông tin bộ dữ liệu**:
   Tên dataset: "IMDB Largest Review".
   Link : https://www.kaggle.com/ebiswas/imdb-review-dataset. [3]
   Thông tin : Bộ dữ liệu có 5571499 review phim ở trang web IMDB dưới dạng 6 file json và chưa được gán nhãn, gồm 9 Feature : (helpful,movie,rating,review date,review detail,review id,review summary,reviewer,spoiler tag).
2. **Tiền xử lí dữ liệu**: Trước khi tiến hành chay mô hình, chúng tôi thực hiện tiền xử lý dữ liệu theo các bước sau đây:

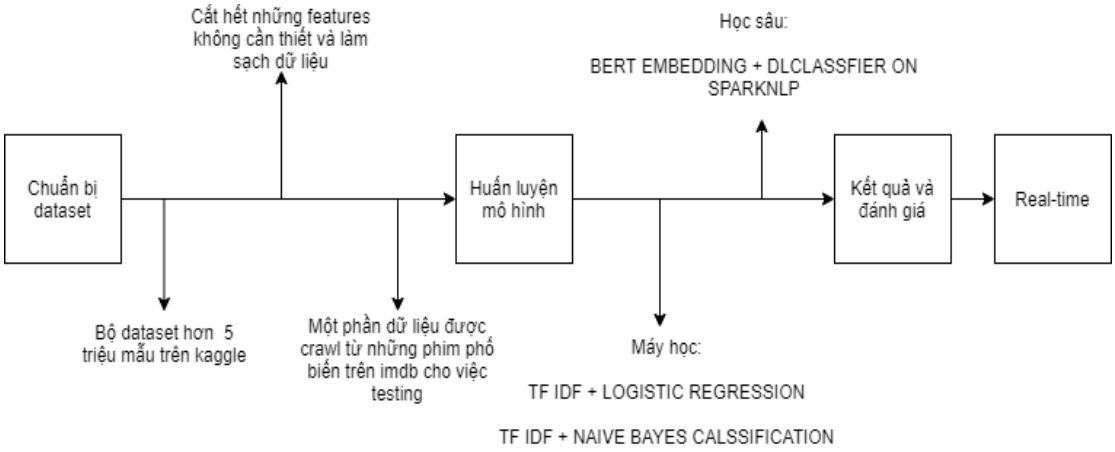
Bảng 1: Các bước tiền xử lý dữ liệu.

Bước	**Cách xử lý**	**Ví dụ**
1	Chuẩn hóa thành text thường	LOKI, THOR -> loki, thor
2	Loại bỏ @Mention khỏi text	this movie is pretty good @John. ->this movie is pretty good.
3	Xóa mã Hashtag, tag, html khỏi text	#thebestfilms -> thebestfilms
4	Loại bỏ URL khỏi text	the sequel is great, it's here https://www.imdb.com/. ->the sequel is great
5	Loại bỏ số cũng như các ký tự không cần thiết	:) ->smile
6	Xóa những khoảng trắng thừa trong câu	this movie so disappointing ->this movie so disappointing.
7	Xóa bỏ những khoảng trắng ở đầu và cuối câu	
8	Loại bỏ các Email	
9	Giữ lại các dòng mà đoạn text có nội dung	

3. **Gán nhãn dữ liệu**: Bộ dữ liệu được gán nhãn dựa trên rating của review về phim:
   * 1-4 : Negative
   * 5,6 : Neutral
   * 7-10 : Positive
4. **Bộ dữ liệu cuối cùng sử dụng trong bài toán**:Vì bộ dữ liệu quá lớn nên chúng tôi chỉ lấy ra 75540 review để sử dụng trong bài toán này và chia làm 2 tập train test với tỉ lệ 80-20. Gồm 2 thuộc tính là (review detail, sentiment).

[3] https://www.kaggle.com/ebiswas/imdb-review-dataset

3.2 Mô tả bài toán



Hình 1: Mô tả chi tiết bài toán.

Input: Những bình luận về phim trên trang web IMDB.com
Output: Phân loại cảm xúc tiêu cực, tích cực của mỗi bình luận theo thời gian thực trên chính trang web IMDB.
Real-time analysising on IMDB: Crawling bình luận và trực tiếp phân loại cảm xúc của bình luận trên trang web IMDB.com. Tổng hợp số lượng bình luận tiêu cực và tích cực của mỗi phim.

3.3 Các mô hình sử dụng

3.3.1 TF-IDF vectorizer kết hợp máy học truyền thống

1. **TF-IDF (Term Frequency – Inverse Document Frequency)** : là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. Tf-idf cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.
   - **TF**:Term Frequency(Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản(tổng số từ trong một văn bản).Như hình 2
   - **IDF**: Inverse Document Frequency(Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ . Khi tính toán TF , tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “is”, “of” và “that” thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.Như hình 3

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

Trong đó:

- tf(t, d): tần suất xuất hiện của từ t trong văn bản d
- f(t, d): Số lần xuất hiện của từ t trong văn bản d
- max({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

Hình 2: TF

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Trong đó:

- idf(t, D): giá trị idf của từ t trong tập văn bản
- |D|: Tổng số văn bản trong tập D
- |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

Hình 3: TF

2. **Logistic Regression(LR):** là một trong những mô hình phân loại cơ bản thường được sử dụng cho các bài toán phân loại nhị phân. Đối với các bài toán phân loại, chúng ta phải chuyển đổi văn bản thành vectơ trước khi lắp chúng vào mô hình đào tạo. Bên cạnh bài toán con sử dụng đầu vào nhị phân như phát hiện loại khía cạnh thì nhiệm vụ phân loại cảm xúc của khía cạnh của chúng tôi có sử dụng bài toán phân loại nhiều lớp nên sẽ có dùng thêm cơ chế one vs res của Logistic Regression.

3. **Naive Bayes (NB):** Ưu điểm của Naïve Bayes (NB) là nó chỉ yêu cầu một lượng nhỏ dữ liệu huấn luyện để ước tính các tham số cần thiết cho việc phân loại. Hơn nữa, NB còn là một mô hình xác suất có điều kiện. Vì vậy trong kỹ thuật của Naïve Bayes, ý tưởng cơ bản để tìm ra xác suất của các categories cho một tài liệu văn bản bởi sử dụng xác suất chung của các từ và categories. Trong bài toán này của chúng tôi, chúng đã sử dụng phương pháp Complement Navie Bayes (CNB). Là adaptation của thuật toán Multinomial Naive Bayes (MNB) đặc biệt thích hợp cho các tập dữ liệu không cân bằng. Cụ thể, CNB sử dụng thống kê từ phần bổ sung của mỗi lớp để tính trọng số của mô hình. Các nhà tạo ra CNB theo kinh nghiệm của họ cho thấy rằng các ước tính tham số cho CNB ổn định hơn so với

các ước lượng cho MNB. Hơn nữa, CNB thường xuyên làm tốt hơn MNB về các nhiệm vụ phân loại văn bản.

3.3.2 Bertembedding + ClassifierDL

Ở bài toán này, chúng tôi sử dụng thư viện SparkNLP do Apache Spark chưa hỗ trợ các công cụ để xây dựng model Deep Learning trên spark.
Spark NLP là một thư viện xử lý ngôn ngữ tự nhiên mã nguồn mở, được xây dựng dựa trên Apache Spark và Spark ML. Nó cung cấp một API dễ dàng tích hợp với ML Pipelines và nó được hỗ trợ thương mại bởi John Snow Labs [4] . Các trình chú thích của Spark NLP sử dụng các thuật toán dựa trên quy tắc, học máy và một số trong số đó là Tensorflow chạy ngầm để cung cấp năng lượng cho các triển khai học sâu cụ thể.

Thư viện bao gồm nhiều tác vụ NLP phổ biến, bao gồm mã hóa, tạo gốc, lemmatization, một phần của gắn thẻ giọng nói, phân tích cảm xúc, kiểm tra chính tả, nhận dạng thực thể được đặt tên, v.v. Tất cả chúng đều được bao gồm dưới dạng mã nguồn mở và có thể được sử dụng bởi các mô hình đào tạo với dữ liệu của bạn. Nó cũng cung cấp các mô hình và đường ống được đào tạo trước, mặc dù chúng được dùng như một cách để bạn cảm nhận về cách hoạt động của thư viện chứ không phải để sử dụng trong sản xuất.

1. **ClassifierDL**: ClassifierDL là một kiến trúc dùng cho bài toán phân loại nhiều lớp. ClassifierDL annotator sử dụng mô hình học sâu (DNN) được xây dựng bên trong TensorFlow và hỗ trợ tối đa 50 lớp.
   Input : Sentences Embedding và output : Category Cụ thể ở đây chúng tôi sử dụng BERT Embedding trong Sentences Embedding

2. **BERT Embedding**:
   [5] [6] [7]

   Về cơ bản, text embedding mã hóa các từ và câu thành các vector có độ dài nhất định giúp cải thiện việc xử lý với dữ liệu dạng văn bản. Những từ cùng xuất hiện trong một ngữ cảnh thì khả năng cao có nghĩa giống nhau.

   Bert (viết tắt của bidirectional encoder representation from transformer) là một pre-training Language model được phát triển bởi đội ngũ Google AI gây nên làn sóng đột phá bởi tính hiệu quả và khả năng ứng dụng trên nhiều lĩnh vực khác nhau của NLP trên kỹ thuật Transfer Learning: Question and Answering,Text classification,..vv.

   Bert sử dụng Transformer là một mô hình attention(attention mechanism) học mối tương quan giữa các từ và một phần của từ trong một văn bản. Transformer có hai phần chính: encoder đọc dữ liệu và decoder đưa ra dự đoán. Ở đây bert chỉ sử dụng encoder

[4] https://www.johnsnowlabs.com/
[5] https://viblo.asia/p/bert-buoc-dot-pha-moi-trong-cong-nghe-xu-ly-ngon-ngu-tu-nhien-cua-google-RnB5pGV7lPG
[6] https://towardsdatascience.com/text-classification-in-spark-nlp-with-bert-and-universal-sentence-encoders-e644d618ca32
[7] https://nef.vn/kt/google-bert/

Khác với các mô hình khác chỉ đọc dữ liệu theo một chiều (directinonal), Encoder đọc toàn bộ dữ liệu trong một lần làm cho bert có thể huấn luyện dữ liệu theo cả hai chiều. Do đó, cải thiện khả năng học được ngữ cảnh của từ tốt hơn thông qua các từ xung quanh nó (trái lẫn phải).

Bert sử dụng hai chiến lược là Masked LM (MLM) và Next Sentence Prediction. Với Masked LM, một lượng dữ liệu trước khi đưa vào Ber (thường là 15 phần trăm từ, không cố định) sẽ được thay thế bởi token [MASK].Khi đó mô hình sẽ dựa vào các từ còn lại như context để dự đoán các token. Lossfuntion chỉ tập trung vào đánh giá các token làm cho Bert hội tụ chậm hơn các directional model khác nhưng mang lại hiệu quả cao hơn, giúp hiểu rõ về ngữ cảnh và ngữ nghĩa tốt hơn.Đối với Next Sentence Prediction, mô hình sử dụng một cặp câu là dữ liệu đầu vào và dự đoán xem câu thứ hai có phải là nối tiếp của câu thứ nhất hay không. Ở đây, một nữa dữ liệu đầu vào là các cặp câu nối tiếp và nữa còn lại là câu thứ hai được chọn ngẫu nhiên từ tập dữ liệu. Nhờ vào hai chiến lược này, bert không chỉ học từ với ngữ cảnh và ngữ nghĩa tốt hơn mà còn có thể học được cả một câu như một từ.

4 Thực nghiệm và kết quả

4.1 Chuẩn bị dữ liệu cho bài toán Real-time

Để chuẩn bị dữ liệu thử cho bài toán phân loại cảm xúc sử dụng mô hình pretrained trên thời gian thực (Realtime), chúng tôi thực hiện lần lượt 2 bước sau:

– Crawl thông tin những bình luận đánh giá về những bộ phim điện ảnh phổ biến từ một trang web phim nổi tiếng là [8] với số lượng review cho phim được chọn là khoảng từ 1000 đến 3000. Cụ thể format của bộ dữ liệu sau khi crawl về có dạng bảng như hình sau 4:

timestamp	user	review_detail
3 April 2021	ThomDerd	I am glad I watch...
26 March 2021	brycetulloch-13604	I liked the actio...
4 April 2021	movieliker1	This is another m...
1 April 2021	kenzibit	Rating basically ...
2 April 2021	alfredsmith	This is a mindles...

Hình 4: Framework bộ dữ liệu realtime.

[8] https://www.imdb.com/?ref=nv_home

– Biến đổi format của bộ dữ liệu vừa crawl về đúng format của đầu vào bài toán đã được thực nghiệm và lưu lại trọng số mô hình, để cuối cùng test được trên thời gian thực, đưa ra ứng dụng thực tế cho bài toán.

4.2 Kết quả

Bảng 2: Kết quả thực nghiệm của mô hình TF-IDF vectorizer + Logistic Regeression Classifier cho bài toán.

Nhãn	**Precision**	**Recall**	**F1- score**	**Support**
Positive	**0.79**	**0.97**	**0.88**	10094
Negative	0.78	0.54	0.66	3002
Neutral	0.39	0.08	0.12	1887
Accuracy			**0.77**	14973
Macro avg	0.65	0.53	0.55	14973
Weight avg	0.73	0.77	0.73	14973

Bảng 3: Kết quả thực nghiệm của mô hình TF-IDF vectorizer + NavieBayes cho bài toán.

Nhãn	**Precision**	**Recall**	**F1- score**	**Support**
Positive	**0.67**	**1.0**	**0.81**	10094
Negative	0.5	0	0	3002
Neutral	0	0	0	1877
Accuracy			**0.67**	14973
Macro avg	0.39	0.33	0.27	14973
Weight avg	0.55	0.67	0.54	14973

Bảng 2,3,4 trình bày kết quả thực nghiệm bài toán phân loại cảm xúc của bình luận phim dựa trên Spark của các mô hình tiếp cận đã được thực hiện qua bước tiền xử lý dữ liệu.

Dựa vào kết quả thu được có thể thấy rằng ở phương pháp máy học truyền thống lại cho kết quả tốt hơn phương pháp học sâu, cụ thể là Logistic Regression là phương pháp đã huấn luyện trước thu được kết quả tốt nhất với 77% Accuracy khi đã được tiền xử lý dữ liệu. Ngoài ra, việc sử dụng các bước tiền xử lý dữ liệu thực sự mang lại hiệu quả đối với các phương pháp được thực hiện khi nó làm tăng Accuracy trên cả ba mô hình LR, SVM và mô hình học sâu ClassifierDLA.

Bảng 4: Kết quả thực nghiệm của mô hình Bertembedding + ClassifierDLApproach cho bài toán.

Nhãn	**Precision**	**Recall**	**F1- score**	**Support**
Positive	**0.93**	**0,76**	**0.84**	12291
Negative	0.51	0,57	0,54	2682
Neutral	0	0	0	0
Accuracy			**0.73**	14973
Macro avg	0.48	0.44	0.46	14973
Weight avg	0.85	0.73	0.78	14973

Mặc khác, tuy với 73% Accuracy nhưng kết quả của mô hình học sâu vẫn được chúng tôi đánh giá cao do hiệu suất của nó mạng lại cho các nhãn phân tích của bài toán với 93% Precision cao nhất và 84% F1-Score trong nhãn Positive. Tuy nhiên, kết quả mô hình Navie Bayes chưa được đánh giá cao, vì kết quả của nó thấp hơn nhiêu so với hai mô hình trên. Vì vậy, chúng tôi vẫn cần thêm thời gian để cải tiến được mô hình này.

Qua các phân tích trên, chúng ta có thể thấy rằng kết quả baseline của chúng tôi chỉ mức tương đối tốt cho bài toán. Cũng thông qua sự phân tích này chúng tôi nhận thấy một số hạn chế trong bộ dữ liệu cũng như quy trình thực hiện bài toán của chúng tôi. Cụ thể, một vài features trong bộ dữ liệu chưa cân bằng như là features rating có sự chênh lệch lớn, với những rating cao chiếm đa số trong bộ dữ liệu, trong khi rating ở mức trung bình, thấp chiếm số lượng rất thấp, khó khăn hơn khi bộ dữ liệu lại có kích thước khá lớn nữa, những ảnh hưởng làm cho việc thực nghiệm bài toán dựa trên mô hình pretrained rất tốn tài nguyên và hiệu suất bài toán giảm đáng kể. Hơn nữa, việc gán nhãn dữ liệu cũng chỉ mức tương đối cho bài toán vì những nghiên cứu cho kiểu bài toán vẫn còn rất ít và chưa phổ biến.

Thống kê kết quả và đánh giá

Tuy vậy, bằng những trọng số đã tối ưu của các mô hình pretrained, bài toán chúng tôi cũng được thực hiện tương đối thành công và có kết quả khả quan để ứng dụng. Để làm làm được việc này, chúng tôi thực hiện thống kê các nhãn dự đoán bình luận phim để đưa ra đánh giá về phim đó. Ví dụ như, kết quả thông kê có số lượng nhãn Positive dự đoán của bình luận nhiều vượt trội so với hai nhãn còn lại thì bộ phim sẽ được khách hành đánh giá cao, đáng xem, hay,..và ngược lại tương tự với trường hợp còn lại

5 Kết luận và hướng phát triển

Crawl được những bộ dữ liệu bình luận về phim nổi tiếng trên trang web phim IMDB Xây dựng thành công được bài toán phân tích cảm xúc của bình luận về những bộ phim dựa trên framework Spark sử dụng cho dữ liệu lớn.

Ứng dụng thực tế thành công cho bài toán. Thống kê số lượng review, cũng như review tích cực tiêu cực của phim -> đề xuất các phim hay, được đánh giá cao. Ngoài ra, nếu một cá nhân có số lượng bình luận negative trong một thời gian nhất định vượt ngưỡng nhất định sẽ bị ban và xóa toàn bộ bình luận

Hướng phát triển:

+Xây dựng lại hệ thống gán nhãn dữ liệu tối ưu hơn cho bài toán.

+ Xây dựng bài toán trên nhiều bộ dữ liệu khác tối ưu hơn hơn về nhãn.

+ Có thể phát triển deploy model cho bài toán.

Acknowledgment

Bảng 5: Bảng phân công công việc

Thành viên	**Công việc**
Hồ Anh Dũng	Thu thập,gán nhãn,tiền xử lí dữ liệu, code, chạy mô hình, làm slide, viết báo cáo
Nguyễn Hữu Trường	Tìm hiểu bài toán, code, chạy mô hình, làm slide, viết báo cáo
Võ Đình Tứ	Tìm hiểu bài toán, crawl dữ liệu realtime, chạy realtime cho bài toán, làm slide, viết báo cáo

Tài liệu

1. Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.
2. Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230, 2017.
3. Minqing Hu and Bing Liu. Mining opinion features in customer reviews. 4(4):755–760, 2004.
4. Gayatree Ganu, Noemie Elhadad, and Amélie Marian. Beyond the stars: improving rating predictions using review text content. 9:1–6, 2009.
5. Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848, 2010.
6. Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. pages 815–824, 2011.
7. Wang and associates. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. pages 115–120, 2012.
8. K.M.O. Cheng and R. Lau. Big data stream analytics for near real-time sentiment analysis. *Journal of Computer and Communications*, pages 189–195, 2015.
9. K.M.O. Cheng and R. Lau. Parallel sentiment analysis with storm. *Transactions on Computer Science and Engineering*, pages 1–6, 2016.