Phân tích cảm xúc trong thời gian thực trên mạng xã hội Twitter

Trần Đình Quyền[1], Hà Phan Diệu Phương[2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Vietnam

Abstract. Bài toán của chúng tôi sử dụng bộ dữ liệu phân tích cảm xúc tiếng Anh gồm một triệu sáu câu tweet đã được thu thập từ Twitter. Chúng tôi tiến hành áp dụng các kỹ thuật tiền xử lý, để đưa bộ dữ liệu sạch, phục vụ cho việc học tốt hơn của model. Với pyspark, chúng tôi sử dụng thư viện Mllib để chạy thực nghiệm các phương pháp máy học, sau đó đánh giá chất lượng của bộ dữ liệu, cũng như hiệu suất của từng mô hình trên bộ dữ liệu. Sau khi đánh giá xem xét về độ chính xác của thuật toán, Logistic Regression đã được chúng tôi lựa chọn để tiến hành dự đoán trong thời gian thực. Cụ thể chúng tôi sử dụng Streaming data, những data được lấy trực tiếp từ Twitter và phân tích cảm xúc liên tục. Kết quả thu được với thuật toán Logistic Regression đạt độ đo Accuracy khoảng 77% cho việc phân tích cảm xúc các câu tweet.

Keywords: Streaming data, Machine learning, Pyspark.

1 Giới thiệu

Trong thời đại bùng nổ công nghệ như hiện nay, mạng xã hội truyền thông ngày càng phát triển, dẫn đến các hành động nói thông thường để thể hiện quan điểm, bày tỏ thái độ đối với một sản phẩm, dịch vụ, tổ chức đã và đang dần chuyển thành các dòng trạng thái, các bình luận bằng văn bản trên mạng xã hội. Từ đó tạo ra một kho dữ liệu khổng lồ giúp các nhà doanh nghiệp có thể tận dụng những đánh giá đó để xây dựng hệ thống như tự động nhận diện cảm xúc, qua đó nhằm xác định và giải quyết mối quan tâm của khách hàng, cải thiện chất lượng sản phẩm, dịch vụ.

Hiểu được giá trị mang lại từ nguồn dữ liệu đó, rất nhiều cộng đồng nghiên cứu trong và ngoài nước quan tâm đến việc khai thác giá trị với mục đích có thể đưa ra những chiến lược, những bước đi mới phục vụ cho thực tế. Hiện nay, tiếng Anh vẫn luôn là một ngôn ngữ được phần lớn công chúng quan tâm và khai thác nhất. Bên cạnh đó có thể thấy đã có rất nhiều bộ dữ liệu tiếng Anh đã được xây dựng. Chính vị vậy điều đó đã góp phần rất lớn cho các nhà nghiên cứu, vì dataset làm nền tảng cho quá trình nghiên cứu của họ. Chúng tôi nhận thấy bộ dataset Sentiment140 với kích thước khổng lồ, gồm một triệu sáu câu tweet, bộ dữ liệu này được các sinh viên trường Standfork xây dựng, việc gán nhãn hoàn toàn diễn ra một cách tự động, không phải gán thủ công

như các bộ dataset khác. Vì vậy kích thước của bộ dữ liệu khá ấn tượng. Thông qua bộ dữ liệu này, chúng tôi mong muốn tận dụng kỹ thuật phân tích dữ liệu lớn trong pyspark để khai phá và phân tích bộ dữ liệu này. Thêm vào đó, chúng tôi quyết định nghiên cứu bài toán này với mong muốn làm tiền đề cho các nghiên cứu sâu hơn khác và đồng thời góp phần thúc đẩy lĩnh vực xử lý ngôn ngữ tự nhiên ngày càng lớn mạnh.

Cấu trúc bài báo được trình bày như sau: Cụ thể trong mục 2, chúng tôi sẽ trình bày các công trình liên quan đến bài toán. Trong mục 3, chúng tôi trình bày chi tiết việc khám phá và áp dụng kỹ thuật tiền xử lý cho bài toán phân tích cảm xúc. Tiếp theo ở mục 4, những phương pháp máy học sử dụng thư viện Mllib trong pyspark để thử nghiệm trên bộ dữ liệu sẽ được trình bày và thông qua kết quả nhận được chúng tôi sẽ tiến hành đánh giá phân tích. Tiếp theo, ở mục 6, chúng tôi trình bày về quá trình Streaming data, dự đoán trong thời gian thực. Cuối cùng mục 7 sẽ là kết luận và kế hoạch nghiên cứu tiếp theo.

2. Công trình nghiên cứu

Những lời nói, những phán xét từ phía dân chúng cho đến nay nhận được mối quan tâm đáng kể trong ngôn ngữ học và nghiên cứu tiếp thị. Kèm theo sự phát triển của mạng xã hội, mọi người không ngại đưa ra các nhận định của cá nhân, kèm theo các cảm xúc chứa đựng trong bình luận đó. Trong lĩnh vực nghiên cứu ngôn ngữ học giao tiếp qua máy tính, Vasquez (2011) [1] đã thực hiện phân tích 100 đánh giá tiêu cực trên TripAdvisor, cho thấy rằng các lời nói trong đó thường đồng thời xảy ra với các hành vi, lời nói bao gồm nhận xét tích cực và tiêu cực, thường đề cập rõ ràng đến các kỳ vọng không được đáp ứng và trực tiếp yêu cầu một đền bù hoặc bồi thường nào đó.

Hiện nay, cộng đồng các nhà nghiên cứu đã có những nghiên cứu về bài toán phân tích cảm xúc như bài báo “Automatically Identifying Complaints in Social Media” [2]. Nghiên cứu này sử dụng Distant Supervision, kết hợp các phương pháp học sâu để đạt được hiệu suất nhận dạng các bình luận phàn nàn tiêu cực trên mạng xã hội khá cao. Bộ dữ liệu được tác giả thu thập từ Twitter bằng ngôn ngữ tiếng Anh với chín lĩnh vực. Pang [3] đã sử dụng máy học có giám sát để phân loại các bài đánh giá phim bằng bộ dữ liệu có sẵn trên kho lưu trữ rec.arts.movies.reviews. Hu and Liu [4] đã khai thác ý kiến trong đánh giá của khách hàng bằng việc tổng hợp các đánh giá của qua việc tóm tắt những ý kiến, khác với tóm tắt văn bản thông thường, họ chỉ quan tâm đến các tính năng cụ thể của sản phẩm và cả về ý kiến tích cực hoặc tiêu cực. Song, phần lớn các nghiên cứu này chưa chú trọng vào việc làm việc với một bộ dữ liệu lớn, chủ yếu tập trung vào việc phân tích cảm xúc mà vẫn còn ít những nghiên cứu áp dụng pyspark vào để phục vụ cho bài toán phân tích cảm xúc.

3. Khám phá và tiền xử lý dữ liệu

3.1 Khám phá dữ liệu

Bộ dữ liệu sentiment140 bằng tiếng gồm một triệu sáu các câu tweet từ người dùng trên Twitter. Dữ liệu bao gồm các tweet về Brand, Product, topic trên mạng xã hội Twitter. Bộ dữ liệu được gán nhãn không qua quá trình thủ công mà hoàn toàn được thực hiện một cách tự động. Bộ dữ liệu gồm hai nhãn positive (tích cực) và nagative (tiêu cực). Điểm đặc biệt ở đây là bộ dữ liệu khá cân bằng, với số lượng câu tích cực và tiêu cực đều đạt 800.000 cây tweet. Điều này sẽ không dẫn đến việc làm khó các mô hình học máy, vì khi được học một nhãn quá nhiều, mô hình sẽ bị thiếu kiến thức huấn luyện về nhãn có số câu ít hơn. Tuy nhiên vì bộ dữ liệu này khá lớn, chúng tôi tin rằng sẽ cung cấp cho mô hình ở nhiều khía cạnh những cảm xúc khác nhau.

Một số câu ví dụ trong bộ dữ liệu được thể hiện ở Bảng 1. Câu thứ nhất thể hiện cảm xúc không được tích cực, không đáp ứng được sự hài lòng của họ, vì vậy cảm xúc không vui vẻ đã được bộc lộ qua lời nói của họ. Câu thứ hai thể hiện thái độ tích cực, thông qua lời nói, cách sử dụng từ ngữ chúng ta có thể nhận thấy cảm xúc tích cực từ họ.

Bảng 1: Ví dụ về các tweet tiêu cực và tích cực trong bộ dữ liệu.

	Nội dung	**Nhãn**
1	Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D	Tiêu cực
2	You're a what!! Never saw that one coming! lmao. Have a good day.	Tích cực

Bộ dữ liệu có nguồn gốc từ Twitter, vì vậy mỗi câu tweet đều mang những đặc điểm chung như: @ và sau đó là tên một người dùng, cũng vì là nơi mạng xã hội, nên các bình luận sẽ chứa những thứ khác như Url, icon, các ký tự đặc biệt được sử dụng một cách thoải mái. Bên cạnh đó độ dài các câu tweet cũng không quá dài, độ dài trung bình bộ dữ liệu khoảng 73 từ cho một câu.

3.2 Tiền xử lý dữ liệu

Sau khi tìm hiểu và khám phá những đặc điểm của dữ liệu, chúng tôi tiến hành áp dụng các kỹ thuật xử lý phù hợp để đưa ra bộ dữ liệu sạch hơn nhằm nâng cao độ chính xác các thuật toán máy học. Chúng tôi chú trọng một vài kỹ thuật phổ biến như xóa đi những URL không cần thiết, các ký tự đặc biệt cũng được chúng tôi loại bỏ. Để dễ dàng xử lý chúng tôi chuyển tất cả chữ viết về chữ thường, xử lý nội dung của các thẻ HTML, xóa đi những hashtag, email, tên người dùng không cần thiết. Sau khi xử lý, chúng tôi tiến hành đưa bộ dữ liệu vào pipeline để phục vụ cho mô hình thuật toán.

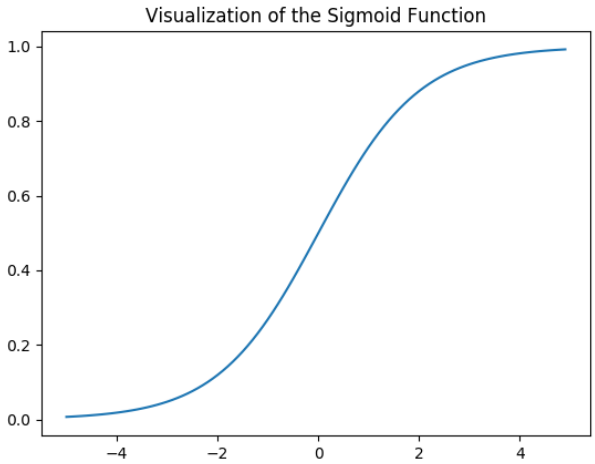
Bảng 2: Ví dụ về các câu đã qua tiền xử lý và chưa tiền xử lý.

	Chưa tiền xử lý	Tiền xử lý
1	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D	awww that s a bummer you shoulda got david carr of third day to do it d
2	|@markhardy1974 Me too #itm	me too itm

4. Các phương pháp tiếp cận

4.1 Logistic Regression

Mô hình này thường được dùng để ước tính xác suất mà một điểm dữ liệu thuộc về một lớp cụ thể. Mô hình này kết hợp mô hình tuyến tính và hàm kích hoạt sigmoid để dự đoán đầu ra. Hàm của mô hình có những tính chất quan trọng. Đây là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1). Trong hình trên, hàm này coi điểm có tung độ là 0.5 là ngưỡng (ngưỡng có thể thay đổi tùy chỉnh), các điểm càng xa ngưỡng về phía bên trái có giá trị càng gần 0, các điểm càng xa ngưỡng về phía bên phải có giá trị càng gần 1 (bài toán phân lớp nhị phân). Vì hàm của mô hình Logistic Regression có đạo hàm mọi nơi, vì vậy có thể được lợi cho việc tối ưu. Ứng dụng được sử dụng trong bài toán phân loại tin nhắn rác, phát hiện gian lận, phát hiện ung thư lành tính hay ác tính,...



Hình 1: Hàm sigmoid trong Logistic Regresstion.

Thuật toán này sau khi áp dụng một vài kỹ thuật nhỏ như one-vs-one, phân tầng, binary coding, one-vs-rest,... thì có thể giải quyết những bài toán phân lớp đa lớp (multi – class classification) cho nhiều bài toán thực tế.

4.2 Naive Bayes

Naïve Bayes là một thuật toán đơn giản nhưng mạnh mẽ đáng ngạc nhiên cho bài toán phân loại nhị phân và đa lớp. Dùng mô hình này đào tạo rất nhanh do chỉ có xác suất của mỗi lớp và và xác suất của mỗi lớp được cung cấp các giá trị đầu vào khác nhau cần phải được tính toán. Không có hệ số cần phải được trang bị bởi các thủ tục tối ưu hóa.

- Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính A1, A2,...,An = {x1,x2,...,xn}.
- Giả sử có m lớp C1, C2,..,Cm. Cho một phần tử dữ liệu X, bộ phân lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân lớp Bayes sẽ dự đoán X thuộc vào lớp Ci nếu và chỉ nếu: P(Ci|X) > P(Cj|X) (1<= i, j <=m, i != j). Giá trị này sẽ tính dựa trên định lý Bayes.
- Để tìm xác suất lớn nhất, ta nhận thấy các giá trị P(X) là giống nhau với mọi lớp nên không cần tính. Do đó ta chỉ cần tìm giá trị lớn nhất của P(X|Ci) * P(Ci). Chú ý rằng P(Ci) được ước lượng bằng |Di|/|D|, trong đó Di là tập các phần tử dữ liệu thuộc lớp Ci. Nếu xác suất tiền nghiệm P(Ci) cũng không xác định được thì ta coi chúng bằng nhau P(C1) = P(C2) = ... = P(Cm), khi đó ta chỉ cần tìm giá trị P(X|Ci) lớn nhất.
- Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toàn P(X|Ci) là rất lớn, dó đó có thể giảm độ phức tạp của thuật toán Naive Bayes giả thiết các thuộc tính độc lập nhau. Khi đó ta có thể tính: P(X|Ci) = P(x1|Ci)...P(xn|Ci).

Sự đơn giản của thuật toán này mang lại hiệu quả trong các bài toán phân loại văn bản, ví dụ bài toán lọc tin nhắn hoặc email rác. Có ba loại phân bố xác suất thường được sử dụng là Gaussian Naive Bayes, Multinomial Naive Bayes và Bernoulli Naive.

4.3 Decision Tree

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary), Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

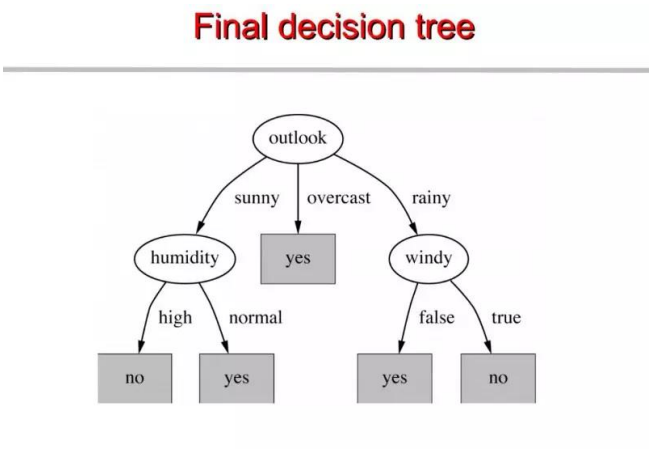
Ưu điểm

- Dễ đọc, dễ hiểu.
- Từ hình ảnh cây đã được tạo dễ dàng phân tích cho ra kết quả, kết quả của nó là tập các luật.
- Có thể giải quyết cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.
- Có thể thẩm định một mô hình các kiểm tra thống kê.

- Có thể xử lý một lượng dữ liệu trong một thời gian ngắn.
- Có thể xử lý một lượng dữ liệu trong một thời gian ngắn.
- Mạnh mẽ đối với các ngoại lệ.

Nhược điểm

- Không phải một thuật toán mang tính phân loại cao.
- Rất dễ bị quá khớp, cần cắt tỉa bớt nhánh, điều chỉnh tham số.
- Không thích hợp với dữ liệu tuyến tính.
- Khó khăn trong việc tính toán khi biến phụ thuộc có nhiều lớp.



Hình 2: Cấu trúc của thuật toán Decision tree.

4.4 TF-IDF

Trong truy hồi thông tin, tf–idf, TF*IDF, hay TFIDF, viết tắt từ cụm từ tiếng Anh: term frequency–inverse document frequency, là một thống kê số học nhằm phản ánh tầm quan trọng của một từ đối với một văn bản trong một tập hợp hay một ngữ liệu văn bản.[1] tf–idf thường dùng dưới dạng là một trọng số trong tìm kiếm truy xuất thông tin, khai thác văn bản, và mô hình hóa người dùng

TF- term frequency – tần số xuất hiện của 1 từ trong 1 văn bản. Cách tính:

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

- Thương của số lần xuất hiện 1 từ trong văn bản và số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản đó. (giá trị sẽ thuộc khoảng [0, 1])
- f(t,d) - số lần xuất hiện từ t trong văn bản d.
- max{f(w,d):w∈d} - số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản.

IDF – inverse document frequency. Tần số nghịch của 1 từ trong tập văn bản (corpus).

Tính IDF để giảm giá trị của những từ phổ biến. Mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản.

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

|D| : - tổng số văn bản trong tập **D**

|{d∈D:t∈d}| : - số văn bản chứa từ nhất định, với điều kiện {\displaystyle t}{\displaystyle t} xuất hiện trong văn bản d (i.e., {\displaystyle \mathrm {tf} (t,d)\neq 0}{\displaystyle \mathrm {tf} (t,d)\neq 0}). Nếu từ đó không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho không không hợp lệ, vì thế người ta thường thay bằng mẫu thức 1+ |{d∈D:t∈d}|

Cơ số logarit trong công thức này không thay đổi giá trị của 1 từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Tuy nhiên việc thay đổi khoảng giá trị sẽ giúp tỷ lệ giữa IDF và TF tương đồng để dùng cho công thức TF-IDF như bên dưới.

Giá trị TF-IDF:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

5. Kết quả thực nghiệm

5.1 Đánh giá hiệu quả tiền xử lý dữ liệu

Sau khi xử lý và loại bỏ các thông tin không cần thiết trong bộ dữ liệu, chúng tôi tiến hành đánh giá độ hiệu quả của việc tiền xử lý. Từ đó hiểu thêm về tầm quan trọng của quá trình này trong quá trình thực hiện bài toán.

Bảng 3: Kết quả đối chiếu khi tiền xử lý và chưa qua tiền xử lý.

	Tiền xử lý	Chưa tiền xử lý
Logistic Regression	77.42%	75%

Qua đó, chúng tôi nhận thấy việc tiền xử lý dữ liệu đã mang lại kết quả tốt hơn cho mô hình, từ đó giúp mô hình có một bộ dữ liệu sạch hơn, dễ dàng học từ đặc điểm quan trọng hơn trong bộ dữ liệu.

5.2 Các mô hình học máy

Sau khi tiền xử lý đã mang lại kết quả tốt hơn cho mô hình, chúng tôi tiến hành đưa dữ liệu đã qua xử lý vào Pipeline để huấn luyện mô hình. Pipeline của chúng tôi xây dựng bao gồm các bước sau:

- Tách từ, vì các câu được cấu thành bởi các từ, cụm từ khác nhau, việc tách từ để đóng vai trò quan trọng.
- Sau khi tách từ, chúng tôi tiến hành loại bỏ những từ không cần thiết (stopwords) những từ này không mang lại giá trị, thậm chí gây nhiễu cho quá trình học của mô hình. Nên chúng tôi tiến hành loại bỏ.
- Tiếp theo, chúng tôi tiến hành vector hóa các dữ liệu chữ, chúng tôi sử dụng TF-IDF để đánh giá tầm quan trọng của từ trong bộ dữ liệu. TF-IDF sẽ loại bỏ những từ không cần thiết, có thể coi như stopwords. Cuối cùng sẽ đưa những vector để phục vụ cho việc đưa vào model huấn luyện.

Qua quá trình thực nghiệm các mô hình máy học của thư viện Mllib trong pyspark, chúng tôi đã thu được kết quả như bảng dưới đây.

Bảng 4. Kết quả thực nghiệm các thuật toán máy học.

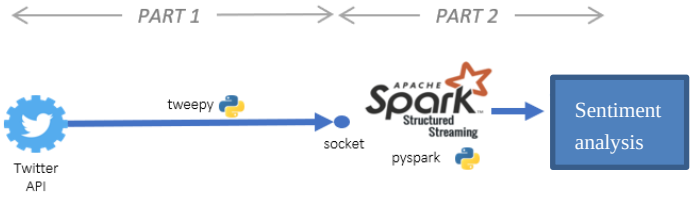
	Accuracy (%)
Naive Bayes	38.14
Decision tree	53.1
Logistic Regression	77.42

6. Streaming data

Ở mục 5, chúng tôi nhận thấy Logistic Regression mang lại kết quả tốt nhất cho việc phân tích cảm xúc trên bộ dữ liệu. Chính vì vậy, chúng tôi quyết định chọn mô hình Logistic Regression để phân tích cảm xúc trong thời gian thực trên Twitter.

Các bước trong quá trình thu thập dữ liệu và xử lý sẽ được thực hiện như dưới đây:

Hình 3: Quá trình streaming data.



Đầu tiên chúng tôi sử dụng Twitter API, sau khi có được thông tin tài khoản để khởi đầu cho việc thu thập dữ liệu. Chúng tôi sử dụng thư viện tweetpy để kết nối và thu thập về các tweet. Sau đó chúng tôi lưu dữ liệu trong thời gian thực vào socket. Từ đó đưa các câu tweet qua giai đoạn tiền xử lý dữ liệu, tiếp theo là đưa câu tweet đã xử lý

qua pipeline dự đoán kết quả. Kết quả thu được sẽ bao gồm một trong hai nhãn là tích cực hoặc tiêu cực. Dự đoán luôn được thực hiện trong thời gian thực.

6.1 Kết nối và thu thập dữ liệu

- Gửi tweet từ API Twitter

Trong phần này, chúng tôi sử dụng thông tin đăng nhập nhà phát triển của mình để xác thực và kết nối với API Twitter. Chúng tôi cũng tạo một TCP socket giữa API của Twitter và Spark, socket này sẽ đợi lệnh gọi của Luồng có cấu trúc Spark và sau đó gửi dữ liệu Twitter. Ở đây, chúng tôi sử dụng thư viện Tweepy của Python để kết nối và nhận các tweet từ API Twitter.

- Tiền xử lý tweet và phân tích cảm xúc

Trong phần này, chúng tôi nhận dữ liệu từ cổng TCP và tiền xử lý bằng thư viện pyspark, là API của Python cho Spark. Sau đó, chúng tôi áp dụng phân tích tình cảm bằng cách sử dụng các mô hình huấn luyện để xử lý dữ liệu dạng văn bản.

6.2 Dự đoán kết quả trong thời gian thực

Các câu tweet đã được đưa qua các bước tiền xử lý để giúp mô hình mang lại kết quả tốt hơn. Thuật toán Logistic Regression sử dụng kiến thức đã được học qua một triệu sáu câu tweet để dự đoán trong thời gian thực. Kết quả thu được như hình dưới đây:

Bảng 5. Kết quả dự đoán

Text	Label
RT What a half of football	4.0
Football track wrestling and soccer	4.0
JusticeForArmyStudents	4.0
If is to retire I don’t know what football or the will look like without him	0.0
Really poor half of football with very little quality to speak of. Wasall though	0.0
…	…
Hey guys	4.0
RT Lets do this CRICKET VS FOOTBALL	4.0
Mint your NFTs cha	4.0
If you have it available	0.0
I was made a believer	4.0

7. Kết luận và hướng phát triển

Trong nghiên cứu này, chúng tôi đã sử dụng bộ dữ liệu gồm một triệu sáu câu tweet để huấn luyện cho mô hình học máy trong thư viện Mllib của pyspark. Bên cạnh đó, chúng tôi đã tìm hiểu và nghiên cứu về Streaming data, áp dụng thuật toán học máy để phân tích cảm xúc trong thời gian thực. Trong quá trình phân tích bộ dữ liệu, chúng tôi nhận thấy vai trò của việc tiền xử lý dữ liệu bài toán đã góp phần cải thiện kết quả bài toán.

Từ kết quả thực nghiệm lần này, đã tạo ra những thách thức lớn khiến chúng tôi phải tiếp tục nghiên cứu sâu hơn, khắc phục những nhược điểm hiện tại và tìm ra những phương pháp mới giúp nâng kết quả cho bài toán phân tích cảm xúc trong tương lai. Những giải pháp có thể được triển khai là cải tiến độ chính xác của bộ dữ liệu, áp dụng các phương pháp học máy khác. Cao hơn nữa là sử dụng các mô hình học sâu để phân tích cảm xúc đạt hiệu quả cao hơn. Ngoài ra chúng tôi cũng cần tìm hiểu thêm về các kỹ thuật trong tiền xử lý dữ liệu trong tiếng Anh để xử lý tốt hơn.

References

1. Preotiuc-Pietro, Daniel, Mihaela Gaman, and Nikolaos Aletras. "Automatically identifying complaints in social media." arXiv preprint arXiv:1906.03890 (2019).
2. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentimentclassification using machine learning techniques.arXiv preprint cs/0205070, 2002.
3. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. InProceed-ings of the tenth ACM SIGKDD international conference on Knowledge discoveryand data mining, pages 168–177, 2004.
4. C. Vásquez, “Complaints online: The case of TripAdvisor,” J. Pragmat., vol. 43, no. 6, pp. 1707–1717, 2011.