APPLE'S TECHNOLOGY PRODUCT REAL-TIME MINING ANALYSING ON TWITTER

Hồ Đình Long[1], Đỗ Hùng Dũng[2], and Nguyễn Thiên Long[3]

[1] KHDL2018, University of Information Technology, HCM, Vietnam
18521022@gm.uit.edu.vn
[2] KHDL2018, University of Information Technology, HCM, Vietnam
18520629@gm.uit.edu.vn
[3] KHDL2018, University of Information Technology, HCM, Vietnam
18521046@gm.uit.edu.vn

Keyword: Apple, Mining, Twitter, Real-time.

1 Giới thiệu

Trước sự phát triển mạnh mẽ của khoa học – công nghệ, xuất hiện càng nhiều các trang mạng xã hội đã tác động lớn đến mọi lĩnh vực hoạt động và sinh hoạt của con người, nhất là giới trẻ.
Hiện nay thế giới có hàng trăm mạng xã hội khác nhau, trong đó một trong mạng xã hội phát triển nhanh nhất và thành công nhất mặc dù có mặt khá muộn, đó là Twitter.
Với lượng người dùng lên đến 500 triệu người, Twitter cho phép người dung chia sẻ các thông tin qua việc đăng các tin nhắn trong phạm vi 140 ký tự, được gọi là tweet. Các nghiên cứu cho thấy các tweet có chứa nhiều loại thông tin, đa dạng về thể loại.

Thu thập và phân tích dữ liệu truyền thông xã hội là bài toàn có tính ứng dụng cao, và ngày càng trở nên thách thức khi mà số lượng dữ liệu đang dần tăng lên một cách chóng mặt, dẫn đến tình trạng quá tải thông tin. Rất nhiều thông tin hữu ích có thể sẽ bị mất do các tweet mới hơn cập nhật và đẩy lùi các tweet trước đó. Từ đó bài toán thu thập và phân tích dữ liệu thời gian thực được quan tâm và sử dụng để khắc phục được khó khăn trên.

Ở bài toán này, chúng tôi tiến hành thu thập và phân tích những tweet có liên quan tới các sản phẩm của nhà sản xuất Apple. Dựa trên những cảm xúc của người dùng ở những tweet đó, từ đó áp dụng các thuật toán Realtime để tìm ra những vấn đề mà người dùng các sản phẩm công nghệ đang quan tâm tại thời điểm đó, cũng như những điểm cần cải thiện trong các sản phẩm để có thể nâng cao trải nghiệm người dùng.

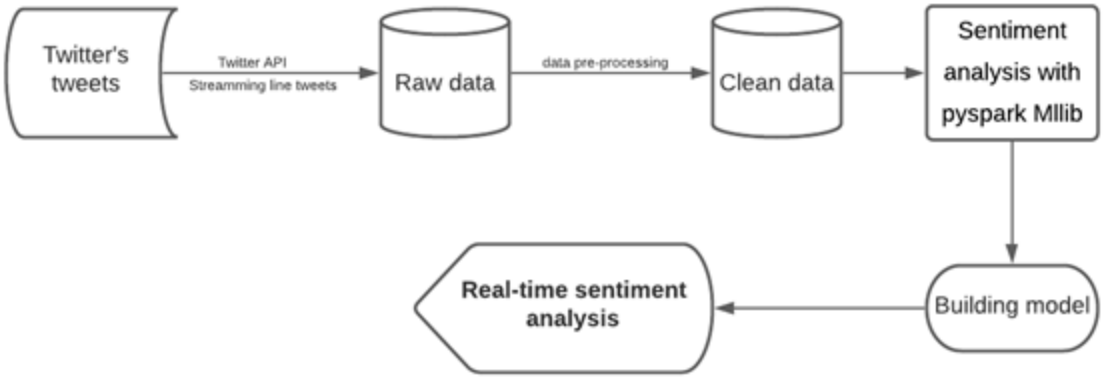
- Input: các tweet về sản phẩm cùa nhà sản xuất Apple trên twitter.
- Output: phân loại cảm xúc được gán 1 trong 3 nhãn : positive, negative, neutral.

Đồ án này được chia làm 7 phần. Phần 1: Giới thiệu. Phần 2: Quy trình thực hiện bài toán. Phần 3: Phân tích bộ dữ liệu sử dụng. Phần 4: Thực nghiệm, bao gồm 2 chỉ mục là cài đặt thử nghiệm và kết quả. Phần 5: Kết luận. Phần 6: Trình bày về những hạn chế mà bài toán này gặp phải. Phần 7: Ứng dụng và tầm nhìn phát triển trong tương lai.

2 Quy trình

Để thực hiện đúng tiêu chí đặt ra của đề tài là cập nhập nhanh chóng, bắt kịp xu hướng tiêu dùng, nhóm chúng tôi tiến hành thu thập các tweets mới nhất có liên quan đến sản phẩm của tập đoàn Apple từ ứng dụng mạng xã hội Twitter bằng công cụ "Twitter API [1]" và phân luồn dữ liệu thu được bằng "spark streaming [2]". Tuy nhiên, để quá trình huấn luyện được diễn ra thuận lợi hơn, chúng tôi có can thiệp vào dữ liệu bằng các phương pháp làm

sạch. Với dữ liệu đã qua quá trình "pre-processing", chúng tôi tiếp tục huấn luyện bằng các mô hình học máy được cung cấp bởi thư viện Pyspark Mllib, kết hợp với pipeline Model [5] cho ra 6 mô hình được huấn luyện từ bộ dữ liệu đã thu thập (sẽ được trình bày ở phần sau). Đến đây, chúng tôi đã có thể sử dụng mô hình đã huấn luyện để chạy "real-time" với các dòng tweets được đăng tải liên tục trên twitter.
Để trực quan quy trình, chúng tôi đã vẽ ra sơ đồ thực hiện, nhằm dễ dàng quan sát tiến độ và tiến hành công việc.



Hình. 1. Quy trình thực hiện bài toán

3 Dữ liệu

Bộ dữ liệu chúng tôi thu thập được 9998 câu bình luận trên Twitter và được gán nhãn cảm xúc: positive ,negative, neutral với mô hình có sẵn của Johnsnowlabs [3].
Bộ dữ liệu của chúng tôi sau khi được gán nhãn bao gồm 4 thuộc tính:
timestamp : thời gian người dung bình luận
user : người dùng Twitter
text : câu bình luận của người dung
user_sentiment : cảm xúc của câu bình luận , 3 nhãn bao gồm:

- Positive: tích cực
- Negative: tiêu cực
- Neutral: trung tính

User_sentiment	count
Positive	6580
Neutral	424
Negative	2294

Hình. 2. Số lượng nhãn cảm xúc trên bộ dữ liệu

Chúng tôi đã liệt kê ra số lượng nhãn cảm xúc trên bộ dữ liệu. Theo như Hình. 2. có thể dễ dàng nhìn thấy số lương nhãn positive chiếm nhiều nhất gấp đôi nhãn nhiều thứ 2 là negative , còn nhãn neutral ít nhất chỉ với 424 nhãn.
Đó là sự mất cân bằng trong bộ dữ liệu mà chúng tôi phải đối mặt. Do đó, chúng tôi muốn cân bằng dữ liệu bằng cách thêm nhãn neutral và negative, cũng như xem xét thủ công từng câu bình luận để nhãn được có độ chính xác cao nhất.

4 Thực nghiệm

4.1 Cài đặt thực nghiệm

Trong bài báo này chúng tôi xây dựng nhiều Pipeline model [5] để thử nghiệm chọn ra được mô hình tốt nhất. Chúng tôi sử dụng feature extraction , feature transformation [6] mặc định trong thư viện Spark MLlib bao gồm TF-IDF, CountVectorizer, HashingTF, Tokenizer , VectorAssembler, Ngram (với n = 3), Stringlndexer kết hợp với 2 mô hình phân là logistic regression [7], random forest [4] . Chúng tôi sử dụng bộ dữ thu thập được với bình luận trong quá khứ để huấn luyện đã được tiền xử lý như xóa các mentions, hastag, URL , email, "RT", chia dữ liệu thành 2 phần là train và test theo tỷ lệ 7:3. Huấn luyện trên bộ train và đánh giá trên bộ test. Sau khi xây dựng được mô hình tốt nhất, chúng tôi sử dụng mô hình để tiến hành phân tích dữ liệu thời gian thực trên Twitter.

4.2 Kết quả

Sau khi tiến hành, chúng tôi thu được 6 kết quả được liệt kê ở Hình. 3. Vì bộ dữ liệu có sự chênh lệch, chúng tôi sử dụng độ đo Accuracy để đánh giá độ chính xác của mô hình.

Pipeline model		Accuracy
HashingTF + IDF	Logistic Regression	0.8062
	Random forest	0.7344
CountVectorizer + IDF	Logistic Regression	0.8075
	Random forest	0.7177
CountVectorizer + Ngram + VectorAssembler	Logistic Regression	0. 8282
	Random forest	0.7466

Hình. 3. Kết quả trên các mô hình đã sử dụng

Từ Hình. 3. ta có thể thấy rằng với mô hình Logistic Regression [7] kết hợp cùng CountVectorizer , Ngram và VectorAssembler cho ra kết quả cao nhất với 82.82%, tuy nhiên chênh lệch không quá nhiều với các sự kết hợp khác của mô hình Logistic Regression. Mô hình Random forest [4] khi kết hợp có kết quả thấp hơn. Nhìn chung, kết quả của các mô hình không quá cao do dữ liệu bị chênh lệch khá nhiều.

5 Kết luận

Qua bài báo này, chúng tôi đã biết được cách tiền xử lý dữ liệu từ twitter, đạt được phân tích dữ liệu sử dụng Apache Spark . Hiểu về cách kỹ thuật extraction và transformation kết hợp với mô hình phân lớp để phân loại cảm xúc. Xây dựng đươc mô hình phân loại cảm xúc đạt kết quả tốt nhất trên mô hình Logistic Regression [7] kết hợp với CountVectorizer , Ngram và VectorAssembler cho kết quả 82.82%.

Ngoài ra, qua thực nghiệm này còn làm nền tảng cho phân tích dữ liệu lớn sử dụng Apache Spark cho sau này. Xây dựng mô hình để phân tích dữ liệu thời gian thực.

6 Hạn chế

Hạn chế về kích thước bộ dữ liệu còn chưa lớn, độ chênh lệch dữ liệu cao, khiến cho mô hình chúng tôi sử dụng chưa ra được kết quả như mong muốn. Một phần vì bộ dữ liệu được gán dựa trên mô hình có sẵn và không có độ chính xác về nhãn quá cao, cũng do chúng tôi không có thời gian để làm thủ công một bộ dữ liệu hoàn chỉnh về bình luận của sản phẩm Apple. Ngoài ra, Twitter còn giới hạn số các bình luận tweet trong phạm vi 140 ký tự, khiến người dùng không thể đăng tải những dòng trạng thái dài hơn. Giả sử bản thân muốn đăng

một điều gì đó lên Twitter, nhưng chỉ có thể đăng một đoạn ngắn với kích thước 70 - 80 ký tự khó có thể bộc lộ được hết cảm xúc tại thời điểm đó, dẫn tới dữ liệu chưa được trọn vẹn.

Trong tương lai, chúng tôi mong muốn có thể cải thiện, tạo một bộ dữ liệu riêng về Apple được gán nhãn thủ công và tìm ra các mô hình có thể phân loại với tỉ lệ chính xác cao hơn. Đồng thời chúng tôi nhắm tới khả năng có thể phát triển trên bộ dữ liệu có quy mô lớn, những bộ dữ liệu về các khía cạnh khác để tìm ra những bài toán có thể áp dụng với thực tế lí tưởng và mang lại những kết quả có thể giúp ích cho đời sống con người và xã hội.

7 Ứng dụng

Với việc phân tích, đánh giá sản phẩm trong môi trường thương mại ở thời gian thực, kết quả nó tạo ra giúp xây dựng tầm nhìn chiến lược cả ngắn hạn và dài hạn cho doanh nghiệp cũng như xác định mục tiêu chính xác, lựa chọn chiến lược và môi trường kinh doanh hiệu quả thông minh. Từ đó trợ giúp rất lớn trong việc hoạch định chiến dịch marketing. Ngoài ra với việc hiểu được cảm xúc người dùng, nhà phát triển còn có thể vận dụng từ đó nâng cao các sản phẩm của họ, giúp các sản phẩm trong tương lai có thể tiếp xúc và phục vụ với nhiều nhóm người hơn và cải thiện độ tin cậy của nhà sản xuất với người dùng.

Tài liệu tham khảo

1. https://developer.twitter.com/en/apps
2. https://spark.apache.org/docs/latest/streaming-programming-guide.html
3. https://nlp.johnsnowlabs.com/2021/01/18/sentimentdl_use_twitter_en.html
4. https://spark.apache.org/docs/latest/ml-classification regression.html#random-forest-classifier
5. https://spark.apache.org/docs/latest/ml-pipeline.html
6. https://spark.apache.org/docs/latest/ml-features.html
7. https://spark.apache.org/docs/latest/ml-classification-regression.html