

# Real-time Twitter Sentiment Analysis

Mai Đức Thuận

Trường Đại Học Công Nghệ Thông Tin, Thành Phố Hồ Chí Minh, Việt Nam  
<https://www.uit.edu.vn/>  
19522316@gm.uit.edu.vn

**Tóm tắt nội dung** Real-time Analytics là một thuật ngữ được dùng nhiều trong ngành công nghệ thông tin, lập trình nói chung. Real-time Analytics được ứng dụng nhiều vào việc tạo ra các phần mềm hỗ trợ hoạt động của doanh nghiệp. Trong bài báo cáo này, nhóm sẽ trình bày mô hình phân tích dữ liệu thời gian thực các câu tweets trên mạng xã hội Twitter với chủ đề Covid-19.

**Keywords:** Real-time Analysis · Twitter Sentiment Analysis · Apache Spark · Apache Kafka · Elasticsearch · Kibana.

## 1 Giới thiệu

Real-time Analytics (Phân tích thời gian thực) được hiểu đơn giản là phân tích dữ liệu ngay khi dữ liệu đó có sẵn. Nói cách khác, người dùng có được thông tin chi tiết hoặc có thể đưa ra kết luận ngay lập tức/rất nhanh sau đó, gần như ngay lập tức khi dữ liệu đi vào hệ thống của họ.

Real-time Analytics cho phép doanh nghiệp phản ứng nhanh trước nhiều tình huống. Họ có thể nắm bắt cơ hội hoặc ngăn chặn các vấn đề trước khi chúng xảy ra. Để so sánh, các phân tích theo kiểu hàng loạt có thể mất hàng giờ hoặc thậm chí vài ngày để cho ra kết quả. Do đó, các ứng dụng phân tích hàng loạt thường chỉ mang lại lợi ích sau khi có sự hiểu biết thực tế (các chỉ số tụt hậu). Thông tin chi tiết từ Real-time Analytics có thể cho phép doanh nghiệp tiết kiệm thời gian và đảm bảo có được dữ liệu chính xác. Để thực sự có ích, các ứng dụng Real-time Analytics phải có tính sẵn sàng cao và thời gian phản hồi thấp nhất có thể.

Ví dụ về Real-time Analytics bao gồm:

- Ghi điểm tín dụng theo thời gian thực, giúp các tổ chức tài chính quyết định ngay lập tức có nên gia hạn tín dụng hay không.
- Quản lý quan hệ khách hàng (CRM), tối đa hóa sự hài lòng và kết quả kinh doanh trong mỗi lần tương tác với khách hàng.
- Phát hiện gian lận tại các điểm bán hàng.
- Nhắm mục tiêu khách hàng cá nhân trong các cửa hàng bán lẻ với các chương trình khuyến mãi và ưu đãi.

## 2 Bộ dữ liệu

Twitter là một dịch vụ mạng xã hội trực tuyến miễn phí cho phép người sử dụng đọc, nhấn và cập nhật các mẫu tin nhỏ gọi là tweets, một dạng tiểu blog. Twitter có hơn 500 triệu tweet được đăng mỗi ngày, trong đó có hơn 330 triệu người hoạt động thường xuyên hàng tháng (theo thống kê cuối năm 2019).

Nền tảng Twitter cung cấp đường truy nhập tới khối dữ liệu đó thông qua Twitter API. Twitter API cho phép lập trình viên truy nhập vào một vài nội dung của Twitter bao gồm các timeline, các cập nhật tweets, và thông tin người dùng, ...

Trong đề án này, Dữ liệu được thu thập từ Twitter API, sau đó đưa vào mô hình dự đoán cảm xúc câu tweet.

Input: Câu Tweet về chủ đề cụ thể.

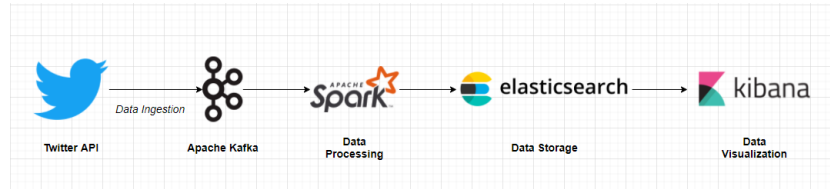
Output:

- Positive.
- Negative.
- Neutral.

## 3 Phương pháp tiếp cận

### 3.1 Tổng quan

Trong mô hình phân tích thời gian thực này, nhóm sẽ lấy dữ liệu thông qua Twitter API dựa trên thư viện Tweepy, tiếp tục đưa dữ liệu thông qua công nghệ Kafka rồi tiền xử lý dữ liệu sử dụng Spark. Dữ liệu được chuyển hóa cuối cùng được đưa vào Elasticsearch và trực quan dữ liệu bằng Kibana. Tổng quan của mô hình như trong Hình 1.



Hình 1. Tổng quan kiến trúc của mô hình.

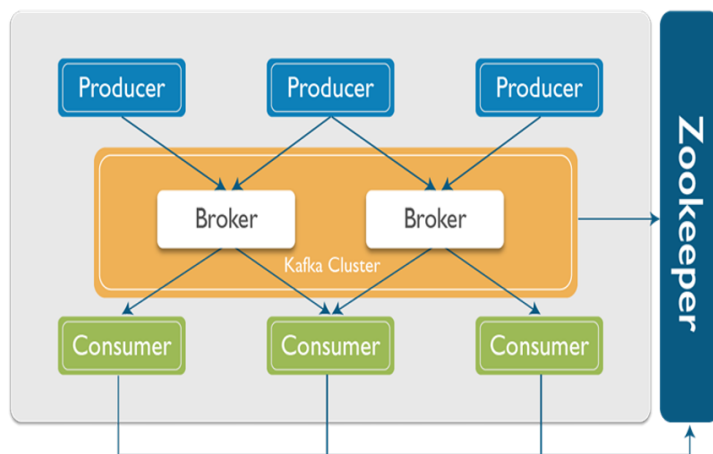
### 3.2 Apache Kafka

Apache Kafka [1] là một nền tảng stream dữ liệu phân tán, là hệ thống message pub/sub phân tán (distributed messaging system). Bên pulbic dữ liệu được gọi là producer, bên subscribe nhận dữ liệu theo topic được gọi là consumer. Kafka có khả năng truyền một lượng lớn message theo thời gian thực, trong trường

hợp bên nhận chưa nhận message vẫn được lưu trữ sao lưu trên một hàng đợi và cả trên ổ đĩa bảo đảm an toàn. Đồng thời nó cũng được replicate trong cluster giúp phòng tránh mất dữ liệu.

Các khái niệm trong Apache Kafka:

- Producer: Một producer có thể là bất kì ứng dụng nào có chức năng publish message vào một topic.
- Messages: Messages đơn thuần là byte array và developer có thể sử dụng chúng để lưu bất kì object với bất kì format nào.
- Consumer: Một consumer có thể là bất kì ứng dụng nào có chức năng subscribe vào một topic và tiêu thụ các tin nhắn.
- Broker: Kafka cluster là một set các server, mỗi một set này được gọi là 1 broker.
- Zookeeper: được dùng để quản lý và bố trí các broker.



**Hình 2.** Cấu trúc của Apache Kafka.

Kafka được xây dựng dựa trên mô hình publish/subscribe, tương tự như bất kỳ hệ thống message nào khác. Các ứng dụng (đóng vai trò là producer) gửi các messages (records) tới một node kafka (broker) và nói rằng những messages này sẽ được xử lý bởi các ứng dụng khác gọi là consumers. Các messages được gửi tới kafka node sẽ được lưu trữ trong một nơi gọi là topic và sau đó consumer có thể subscribe tới topic đó và lắng nghe những messages này. Messages có thể là bất cứ thông tin gì như giá trị cảm biến, hành động người dùng,...

### 3.3 Tiền xử lý dữ liệu

Dữ liệu được thu thập từ Kafka sẽ được tiền xử lý dữ liệu bằng apache spark [2].

```
{
  'country_iso2': 'US', 'geo': {'location': '37.1232245,-78.4927721'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 13, 693)}
{'uid': '1421143583567060992', 'text': '@DrTessaI @DrNoMask eye surgeon wanting to be relevant to cash in on covid', 'user': 'Lady D153886', 'sentiment': 'Positive', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '39.7837384,-108.4458825'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 869119)}
{'uid': '1421143583768428352', 'text': 'RT @atbwalshBlog: 100 thousand people can die of the flu and we don't bat an eye. No measures are taken. Nothing is done. It's barely disc-', 'user': 'huckic21', 'sentiment': 'Positive', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '48.7127281,-74.0068152'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 29, 767060)}
{'uid': '1421143583831302146', 'text': 'RT @DrMarkSchlissel: As part of our commitment to a safe and vibrant Fall semester, today our Ann Arbor, @UMFlint, @UMDearborn & @UMichMed.', 'user': 'emmabausch', 'sentiment': 'Positive', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '41.8755616,-87.6244212'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 32, 618884)}
{'uid': '1421143584313593859', 'text': 'Canada Officially Welcomes Covid 19's Fourth 'Delta Wave' https://t.co/2aIGk5Schy', 'user': 'ItsDeanlundell', 'sentiment': 'Neutral', 'country': 'Canada', 'country_iso2': 'CA', 'geo': {'location': '43.6534817,-79.3839347'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 45, 866243)}
{'uid': '1421143584728887301', 'text': 'McNab says Freeman is seeing on average one death per day in Joplin due to Covid.', 'user': 'MorningsonKRP5', 'sentiment': 'Negative', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '37.1435741,-94.4634702'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 42, 703850)}
{'uid': '1421143584636641281', 'text': 'RT @LozzaFox: They are called women, you misogynistic twats. https://t.co/Fkisc1Valz', 'user': 'psibergal', 'sentiment': 'Neutral', 'country': 'United Kingdom', 'country_iso2': 'GB', 'geo': {'location': '51.546855800000005,-0.2537791389340719'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 51, 278174)}
{'uid': '1421143585013993476', 'text': '@TheFirstonTV Cause you can't get Covid during a photo opp! 🤔', 'user': 'saraags82', 'sentiment': 'Neutral', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '31.8160381,-99.5120986'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 52, 257635)}
{'uid': '1421143585416781825', 'text': 'RT @RobDenBleyker: angry parents: "wearing masks is traumatizing our kids!"\n\ngovernors: "hm good point, what if instead we let some of thei-', 'user': 'jeff stacey', 'sentiment': 'Positive', 'country': 'Canada', 'country_iso2': 'CA', 'geo': {'location': '49.8955367,-97.1384584'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 55, 305212)}
{'uid': '1421143585601187841', 'text': 'when will covid be done with', 'user': 'carlitoguey508', 'sentiment': 'Neutral', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '37.7884969,-122.3558473'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 20, 57, 83949)}
{'uid': '1421143585924255747', 'text': 'RT @OliviaRoyce: There is no excuse for your ongoing negligent behavior. As Republican officials-@houseGOP-you know better. Your actions ar-', 'user': 'AndJohnson1', 'sentiment': 'Positive', 'country': 'United States', 'country_iso2': 'US', 'geo': {'location': '43.4849133,-71.6553992'}, '@timestamp': datetime.datetime(2021, 7, 30, 16, 21, 0, 575163)}
```

Hình 3. Dữ liệu từ Kafka.

Các bước xử lý dữ liệu:

- Xử lý datetime.
- Loại bỏ ký tự đặc biệt.
- Loại bỏ Hash-tag, mention.
- Chuyển hóa text.

Mô hình trên sẽ sử dụng Textblob để dự đoán cảm xúc của câu tweet đó. TextBlob là thư viện nổi bật để xử lý dữ liệu văn bản, cung cấp một API đơn giản để thực thi các tác vụ NLP như đánh dấu POS, rút trích cụm danh từ, phân tích cảm xúc, phân loại (Naive Bayes, Decision Tree), diễn dịch, tích hợp WordNet, parsing, ... Cuối cùng, dữ liệu được xử lý và đưa vào các bước tiếp theo sẽ trông như Hình 4.

|   | Date                | User                   | IsVerified | Tweet   | Likes | RT | User_location              | clean_tweet                                       | Sentiment |
|---|---------------------|------------------------|------------|---|-------|----|----------------------------|---|-----------|
| 0 | 2020-11-03 18:15:38 | Jem Packer             | False      | @AdamPaker @IanWright0 Cheers mate! But don't ... | 0     | 0  | London                     | cheers mate but don t look at my goalscoring r... | Neutral   |
| 1 | 2020-11-03 18:15:37 | RetiredSubGuy          | False      | @JossSheldon I always bitterly regretted not b... | 0     | 0  |                            | i always bitterly regretted not being able to ... | Positive  |
| 2 | 2020-11-03 18:15:34 | Davy Boyd              | False      | @bradleyjames22 @DB_HighburyAFC He's a twat, w... | 0     | 0  |                            | highburyafc he s a twat who ain t arsenal mate... | Neutral   |
| 3 | 2020-11-03 18:15:23 | Chieftain 🍷 #ENDSARSng | False      | @safiyalm @ShmuckFC @1amees @Yuse_h @astriluc...  | 0     | 0  | Top of the food chain      | h involving arsenal again                         | Neutral   |
| 4 | 2020-11-03 18:15:21 | floof                  | False      | I got bad at arsenal my aim got worse             | 0     | 0  | The end of the rainbow 🌈 🌈 | i got bad at arsenal my aim got worse             | Negative  |

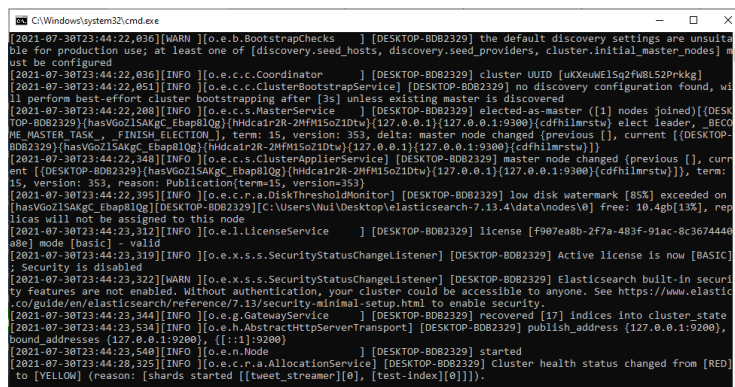
Hình 4. Tiền xử lý dữ liệu..

### 3.4 Elasticsearch và Kibana.

Elasticsearch [3] là một database server mã nguồn mở, độc lập được phát triển bằng Java. Về cơ bản, nó được sử dụng để tìm kiếm và phân tích văn bản (full-text-search).

Nó lấy dữ liệu không cấu trúc từ nhiều nguồn khác nhau và lưu trữ nó ở định dạng phức tạp được tối ưu hóa cao cho các tìm kiếm dựa trên ngôn ngữ. Elasticsearch thực chất hoạt động như 1 web server, có khả năng tìm kiếm nhanh chóng thông qua giao thức RESTful.

Kibana [4] là một nền tảng phân tích và trực quan hóa nguồn mở được thiết kế để hoạt động kết hợp chặt chẽ với Elasticsearch. Kibana cung cấp các tính năng cho người dùng quản lý như biểu đồ cột, biểu đồ đường, biểu đồ tròn, biểu đồ nhiệt và nhiều loại chart khác.



```

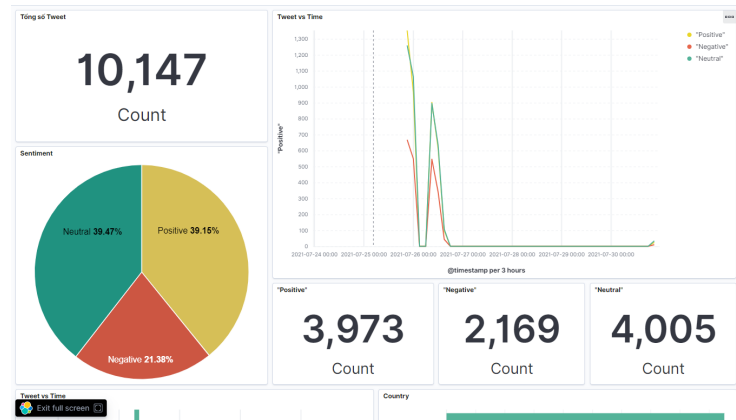
C:\Windows\system32\cmd.exe
[2021-07-30T23:44:22,036][WARN ][o.e.b.BootstrapChecks ] [DESKTOP-BDB2329] the default discovery settings are unsuitable for production use; at least one of [discovery.seed_hosts, discovery.seed_providers, cluster.initial_master_nodes] must be configured
[2021-07-30T23:44:22,036][INFO ][o.e.c.c.Coordinator ] [DESKTOP-BDB2329] cluster UUID [uKXeuNEl5q2FWBL52Prkkg]
[2021-07-30T23:44:22,051][INFO ][o.e.c.c.ClusterBootstrapService] [DESKTOP-BDB2329] no discovery configuration found, will perform best-effort cluster bootstrapping after [3s] unless existing master is discovered
[2021-07-30T23:44:22,208][INFO ][o.e.c.s.MasterService ] [DESKTOP-BDB2329] elected-as-master ([1] nodes joined){[DESKTOP-BDB2329]{hasVGoz15AKgc_Ebap81Qg}{hdcair2R-2MFM15oz1Dtw}{127.0.0.1}{127.0.0.1:9300}{cdfhilmrstw} elect leader, _BECDHE MASTER TASK, _FINISH [tset10w]], term: 15, version: 353, delta: master node changed (previous [], current {[DESKTOP-BDB2329]{hasVGoz15AKgc_Ebap81Qg}{hdcair2R-2MFM15oz1Dtw}{127.0.0.1}{127.0.0.1:9300}{cdfhilmrstw}})}
[2021-07-30T23:44:22,348][INFO ][o.e.c.s.ClusterApplierService] [DESKTOP-BDB2329] master node changed (previous [], current {[DESKTOP-BDB2329]{hasVGoz15AKgc_Ebap81Qg}{hdcair2R-2MFM15oz1Dtw}{127.0.0.1}{127.0.0.1:9300}{cdfhilmrstw}}), term: 15, version: 353, reason: Publication[term=15, version=353]
[2021-07-30T23:44:22,395][INFO ][o.e.c.r.a.DiskThresholdMonitor] [DESKTOP-BDB2329] low disk watermark [85%] exceeded on [hasVGoz15AKgc_Ebap81Qg][DESKTOP-BDB2329][C:\Users\Nui\Desktop\elasticsearch-7.13.4\data\nodes\0] free: 10.4gb[13%], replication will not be assigned to this node
[2021-07-30T23:44:23,312][INFO ][o.e.l.licenseService ] [DESKTOP-BDB2329] license [f907ea8b-2f7a-483f-91ac-8c3674440a8e] mode [basic] - valid
[2021-07-30T23:44:23,319][INFO ][o.e.x.s.s.SecurityStatusChangeListener] [DESKTOP-BDB2329] Active license is now [BASIC]; Security is disabled
[2021-07-30T23:44:23,322][WARN ][o.e.x.s.s.SecurityStatusChangeListener] [DESKTOP-BDB2329] Elasticsearch built-in security features are not enabled. Without authentication, your cluster could be accessible to anyone. See https://www.elastic.co/guide/en/elasticsearch/reference/7.13/security-minimal-setup.html to enable security.
[2021-07-30T23:44:23,344][INFO ][o.e.g.GatewayService ] [DESKTOP-BDB2329] recovered [17] indices into cluster_state
[2021-07-30T23:44:23,524][INFO ][o.e.h.AbstractHttpServerTransport] [DESKTOP-BDB2329] publish_address {127.0.0.1:9200}, bound addresses {127.0.0.1:9200}, {[::1]:9200}
[2021-07-30T23:44:23,540][INFO ][o.e.n.Node ] [DESKTOP-BDB2329] started
[2021-07-30T23:44:28,325][INFO ][o.e.c.r.a.AllocationService] [DESKTOP-BDB2329] Cluster health status changed from [RED] to [YELLOW] (reason: [shards started [[tweet_stream][0], [test-index][0]]]).

```

Hình 5. Khởi chạy Elasticsearch.

## 4 Tổng kết

Trong đồ án này nhóm đã tiến hành thiết kế và xây dựng một mô hình phân tích dữ liệu thời gian thực sử dụng các công nghệ dữ liệu lớn và xử lý thời gian thực như Spark, Kafka, ... để giải quyết bài toán. Kết quả trực quan hóa được thể hiện trong Hình 6.



Hình 6. Trực quan dữ liệu.

Trong tương lai, nhóm sẽ tiếp tục phân tích và có thể tìm ra các thông tin giá trị từ bộ dữ liệu này nhằm cải thiện hiệu suất mô hình. Đồng thời thực nghiệm với các bộ dữ liệu khác nhau, các mô hình khác nhau để tối đa hóa hiệu suất.

## Lời cảm ơn

Em muốn gửi lời cảm ơn đến các Thầy Cô Giảng Viên Trường Đại Học Công Nghệ Thông Tin và đặc biệt là thầy Đỗ Trọng Hợp, người đã tận tình hướng dẫn, chỉ bảo chúng em trong thời gian qua, đồng thời cũng là người đã giúp nhóm hoàn thiện bài báo cáo này.

Em xin chân thành cảm ơn!

## Tài liệu

1. Apache Kafka, 2021, <https://kafka.apache.org/documentation/>
2. Apache Spark, 2021, <https://spark.apache.org/docs/latest/>
3. Elasticsearch, 2021, <https://www.elastic.co/guide/index.html>
4. Kibana, 2021, <https://www.elastic.co/guide/en/kibana/7.13/index.html>