Sentiment Analysis Using Spark And Kafka

Lê Nguyên Hoàng[1], Trần Văn Bảo[1], Trần Lê Duy Anh[2], Lê Thành Công[1], and Lê Thành Danh[1]

[1]FPT Telecom, Ho Chi Minh city, Vietnam
[2]Can Tho University of Technology, Can Tho city, Vietnam

Tóm tắt
Với sự phát triển mạnh mẽ của công nghệ thông tin, các trang mạng xã hội như Facebook cũng trở thành phương tiện hữu ích trong việc kinh doanh trực tuyến, giúp khách hàng thực hiện mua hàng, chia sẻ những trải nghiệm và đánh giá qua các bình luận sau giao dịch. Chính vì thế việc nghiên cứu đề xuất phương pháp khai thác ý kiến và phân tích cảm xúc khách hàng thông qua việc thu thập tập dữ liệu là ý kiến bình luận của khách hàng trên Facebook là vô cùng cần thiết. Sau khi thu thập được dữ liệu, tiến hành huấn luyện NLP tiếng việt với ngữ cảnh là các câu bình luận của khách hàng sau khi mua sản phẩm, sau đó ứng dụng vào một trang Facebook để phân tích cảm xúc, và thống kê các trọng tâm được nói đến nhiều nhất. Kết quả thực nghiệm cho thấy độ chính xác 94% của phương pháp đề xuất và kết quả khai thác được tập thông tin, tri thức tiềm ẩn có giá trị từ tập ngữ liệu nhằm giúp các cửa hàng, doanh nghiệp hiểu được các ưu nhược điểm về sản phẩm, dịch vụ để cải thiện chiến lược kinh doanh tốt hơn.

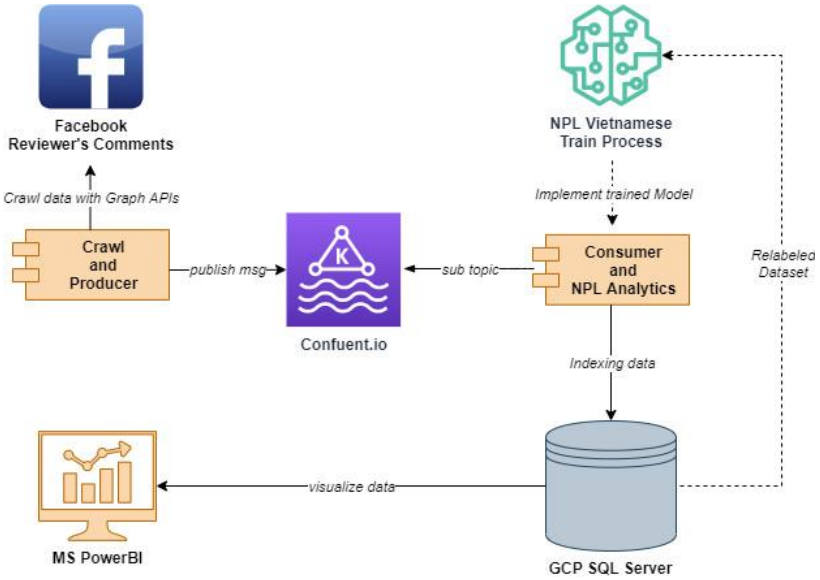
Từ khóa
Facebook, kinh doanh, dữ liệu, bình luận, phân tích cảm xúc, khách hàng, NLP.

1. Giới thiệu

Ngày nay, kinh doanh trực tuyến đã trở thành xu hướng của thời đại, các trang mạng xã hội góp phần không nhỏ trong thị trường kinh doanh trực tuyến, việc bán hàng trên Facebook càng ngày trở nên phát triển và đơn giản mà hiệu quả lại rất cao, mang về nhiều lợi nhuận cho người kinh doanh. Cũng từ đó trên các trang mạng xã hội được nhiều người dùng truy cập sẽ có rất nhiều dữ liệu là các bình luận, đánh giá về sản phẩm của khách hàng. Nhất là những trang của các thương hiệu nổi tiếng, nhãn hàng lớn rất được khách hàng chú ý. Ý kiến khách hàng là những phản hồi mà khách hàng cảm nhận được sau khi sử dụng dịch vụ, hay mua sản phẩm của doanh nghiệp. Những ý kiến của khách hàng có thể tiêu cực hoặc tích cực. Dựa theo những nhận xét tích cực của khách hàng, doanh nghiệp sẽ biết được những ưu điểm của sản phẩm hay dịch vụ mình đang cung cấp. Những ý kiến của khách hàng ngoài việc giúp doanh nghiệp nhìn nhận thực tế về chất lượng dịch vụ cũng như sản phẩm mình đang kinh doanh. Bên cạnh đó có thể dùng những bình luận tích cực để quảng bá hay truyền thông. Từ đó các doanh nghiệp sẽ luôn luôn cải thiện chất lượng dịch vụ, sản phẩm để có thể phát triển và dẫn đầu. Do đó, nhóm chúng em đã chọn đề tài này để phân tích cảm xúc của khách hàng, xem bình luận nào là tích cực, bình luận nào là tiêu cực và vấn đề trọng tâm khách hàng quan tâm là gì. Kết quả của đề tài khi ứng dụng vào trang kinh doanh trực tuyến trên facebook sẽ giúp các cửa hàng, doanh nghiệp nắm bắt thông tin một các dễ dàng và nhanh chóng, từ đó việc phát triển kinh doanh được cải thiện và nâng cao, chẳng hạn việc nâng cao sự hài lòng của khách hàng và giữ chân khách hàng tốt hơn

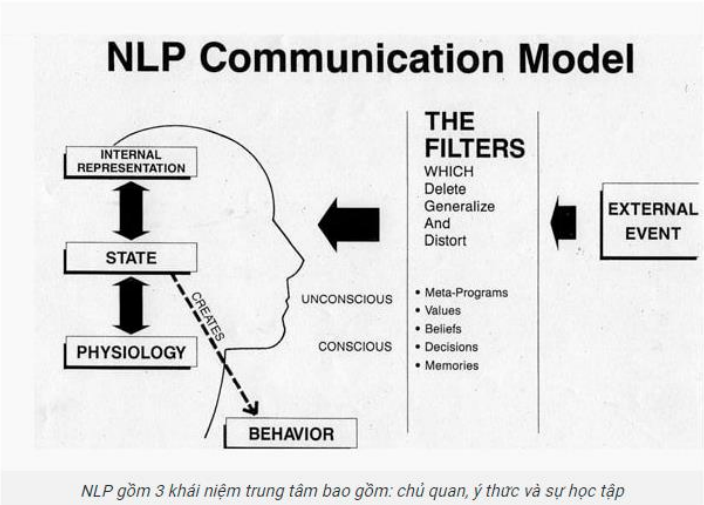
2. System Design

System Design của nhóm bao gồm 7 thành phần cơ bản



2.1 . NLP

NLP (Neuro-linguistic programming) là kỹ thuật coaching ứng dụng tâm lý học với các chiến lược và kỹ thuật hợp lý để mang lại kết quả mong muốn. Phương pháp này dựa trên Neuro (thần kinh học) và Linguistic (ngôn ngữ) của não bộ để "lập trình" (programming) lại tư duy coachee – từ đó mang lại những thay đổi cơ bản trong thái độ, hành vi và cuộc sống.



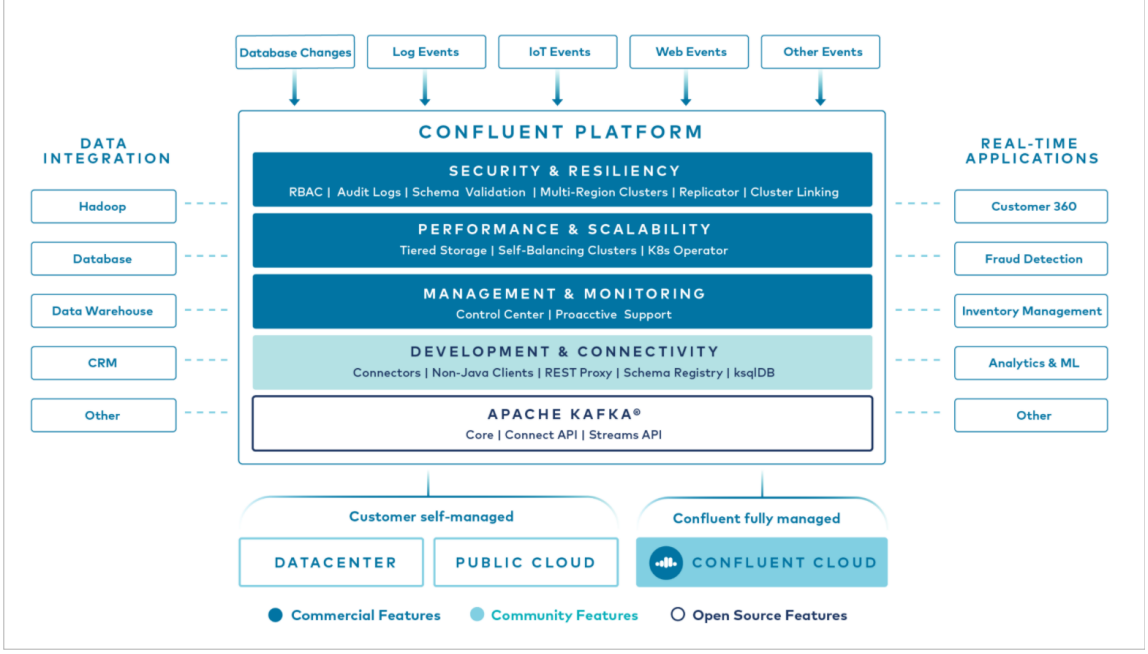
NLP gồm 3 khái niệm trung tâm bao gồm: chủ quan, ý thức và sự học tập

Chủ quan: Chủ quan trong phương pháp lập trình ngôn ngữ tư duy chính là kinh nghiệm. Theo Bandler và Gri, bộ não của chúng ta chính là một thế giới thu nhỏ, thế này sẽ tạo ra chủ quan của chúng ta. Việc chúng ta sinh sống và làm việc với chủ quan này sẽ tạo nên kinh nghiệm. Kinh nghiệm của con người có được từ việc tương tác với thế giới thông qua ngôn ngữ và 5 giác quan. Chính vì tầm quan trọng của chủ quan mà NLP tập trung vào nghiên cứu vấn đề này.

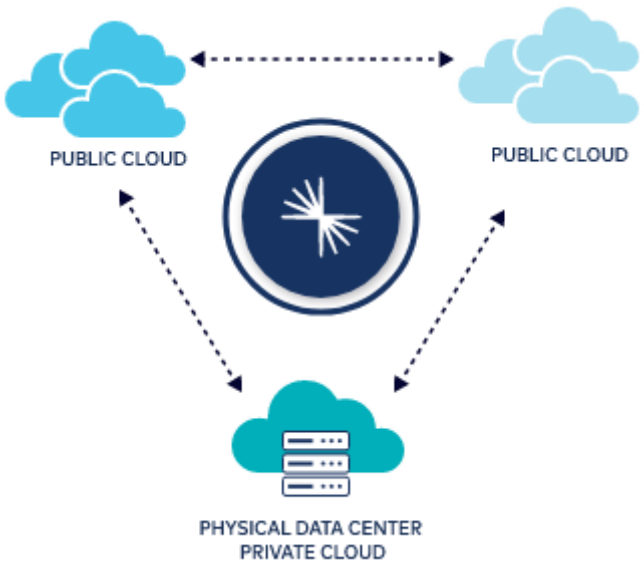
Ý thức: Phương pháp NLP chia ý thức thành 2 phần là ý thức và vô thức. Những trải nghiệm chủ quan mà một người không nhận thức được sẽ gọi là "vô thức".
Học tập: Lập trình ngôn ngữ tư duy NLP áp dụng nguyên lý bắt chước trong mô hình học tập. Việc bắt chước các chuyên gia, thiên tài sẽ giúp bạn học được những kỹ năng quý giá.

2.2. Confluent

Confluent được tạo ra bởi những người sáng lập ra Apache Kafka. Confluent Platform giúp chúng ta xây dựng các ứng dụng streaming một cách đơn giản bằng cách tích hợp dữ liệu từ rất nhiều nguồn tới một nơi tập trung đó là Kafka.



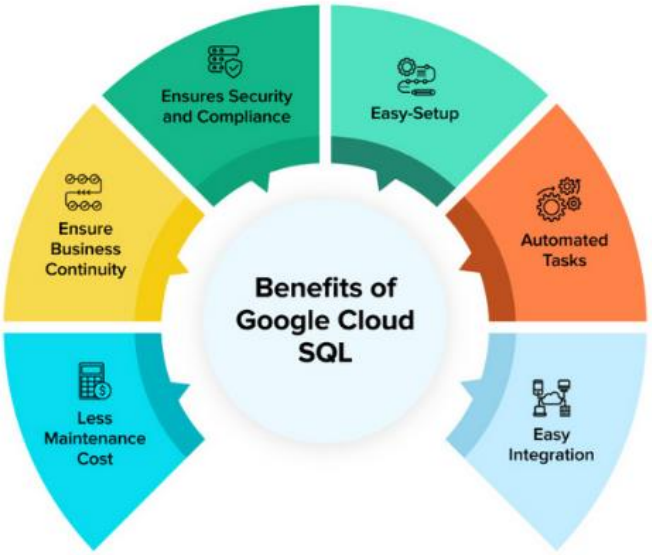
Confluent Platform giúp cho việc xây dựng các data pipeline thời gian thực và các ứng dụng streaming trở nên dễ dàng hơn bằng cách tổng hợp dữ liệu từ nhiều nguồn, tại nhiều vị trí vào một platform streaming dữ liệu duy nhất. Confluent Platform cho phép bạn tập trung vào các logic nghiệp vụ hơn, thay vì lo lắng về các cơ chế cơ bản như cách dữ liệu được vận chuyển hoặc sự tương tác giữa các hệ thống khác nhau. Cụ thể, Confluent Platform đơn giản hóa việc kết nối các nguồn dữ liệu với Kafka, xây dựng các ứng dụng với Kafka, cũng như bảo mật, giám sát và quản lý cơ sở hạ tầng Kafka.



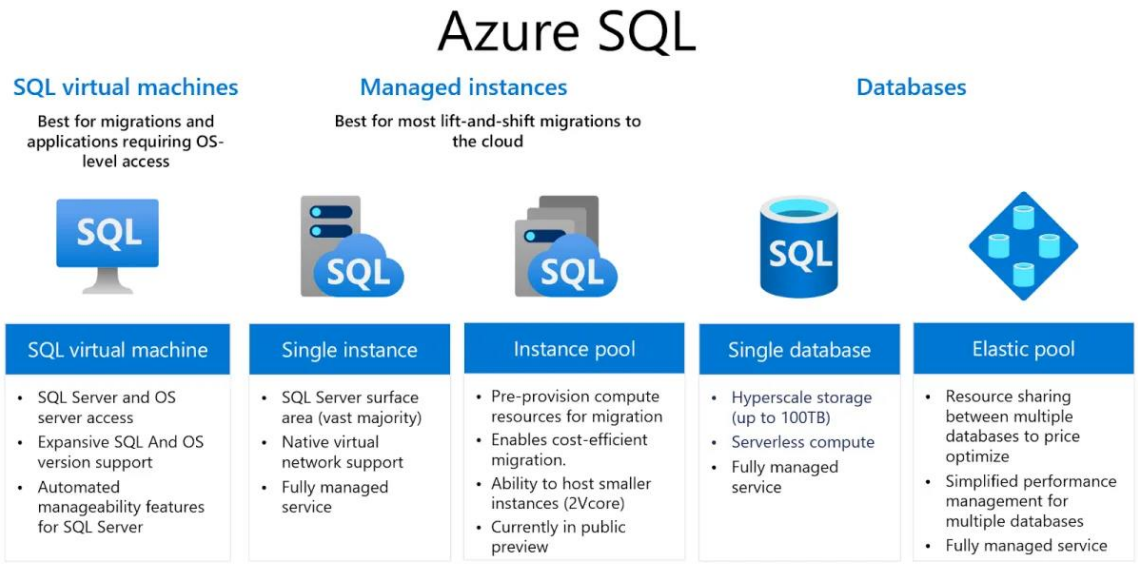
Chuyển các ứng dụng lên đám mây là một công việc vô cùng phức tạp. Và bởi vì một số ứng dụng có thể được liên kết phức tạp với các hệ thống kế thừa, nên việc di chuyển thậm chí có thể không phải là một tùy chọn.
Với Confluent, nắm bắt đám mây theo tốc độ và duy trì một cầu nối dữ liệu liên tục để giữ cho dữ liệu trên tất cả các môi trường tại chỗ, kết hợp và đa đám mây được đồng bộ hóa.Cho phép các nhà phát triển truy cập vào các công cụ đám mây tốt nhất và xây dựng các ứng dụng thế hệ tiếp theo nhanh hơn.

2.3. GCP SQL Server

Cloud SQL cho SQL Server là dịch vụ cơ sở dữ liệu được quản lý hoàn toàn với SLA 99,95%. Được quản lý đầy đủ bao gồm nâng cấp, vá lỗi, bảo trì, sao lưu và điều chỉnh. Tính khả dụng trong khu vực và nhiều hình dạng VM khác nhau với bộ nhớ từ 3,75 GB đến 416 GB và bộ nhớ lên đến 30 TB cho tất cả khối lượng công việc cung cấp các tùy chọn mở rộng linh hoạt để loại bỏ nhu cầu cung cấp trước hoặc lập kế hoạch dung lượng trước khi bắt đầu



Khi chạy SQL Server trên Compute Engine, các máy ảo có thể di chuyển trực tiếp giữa các hệ thống máy chủ mà không cần khởi động lại, điều này giúp các ứng dụng luôn chạy ngay cả khi hệ thống máy chủ yêu cầu bảo trì.



3. Thực nghiệm

Hiện nay, tính đến tháng 06/2020. Tại Việt Nam đã có hơn 69 triệu tài khoản Facebook. Chiếm 2/3 dân số Việt Nam (96,2 triệu người – số liệu năm 2019, Theo gso.gov.vn) Dữ liệu có trọng tâm. Có các nhóm để bình luận về một chủ đề đặc thù. Khả năng tích hợp hệ thống. Facebook hỗ trợ API thuận tiện trong việc thu thập các bình luận.

3.1 API Facebook

Đi sâu hơn về khả năng tích hợp hệ thống, facebook hỗ trợ GraphAPI được đặt tên theo ý tưởng "đồ thị xã hội" - đại diện cho các thông tin trên Facebook. Nó bao gồm:

• Nodes (nút): là các đối tượng riêng như là người dùng, ảnh, trang cá nhân, bình luận…

• Edges (cạnh): là các kết nối giữa những đối tượng riêng ở trên, ví dụ như kết nối hình ảnh và trang chứa hình ảnh đó, bình luận và bức ảnh được bình luận...

• Fields (trường): dữ liệu của đối tượng riêng ở trên, ví dụ như tên, ngày sinh của người dùng, tên trang...

Để sử dụng được facebook api thì cần các thông tin sau:

URL lưu trữ, hầu như tất cả các yêu cầu đều được chuyển đến URL lưu trữ graph.facebook.com. Chỉ có video tải lên sử dụng graph-video.facebook.com..

Access-token: mã xác thực là chủ của một ứng dụng của một tài khoản facebook. Có 3 loại mã: Mã truy cập người dùng: dùng để thay mặt một người sửa đổi hoặc ghi dữ liệu Facebook của người đó. Mã truy cập ứng dụng: dùng để đăng hành động trong Open Graph. Mã truy cập trang: dùng để sửa đổi dữ liệu thuộc về 1 trang Facebook. Có 2 loại thời hạn cho mã là short-term (1-2h) và long-term (60 ngày).

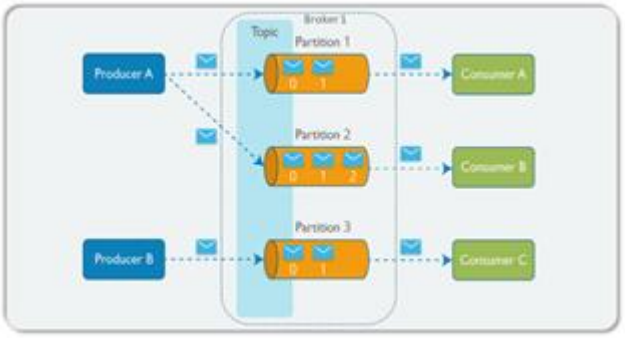
Thực hiện:

- Quét toàn bộ topic trong một trang facebook
- Với mỗi topic lấy toàn bộ bình luận
- Ghi lại mốc thời gian để lần chạy sau chỉ lấy các bình luận mới.

3.2 KAFKA

Vì thông tin crawl được từ mạng xã hội facebook là rất lớn cộng với việc thông tin được đưa về liên tục, nên phải có Distributed Messaging System chịu trách nhiệm ghi nhận và phân phối các thông tin này. Nhóm Kafka – Confluent.io là giải pháp khi thực hiện đề tài Social Listening có phân tích NLP. Kiến trúc Kafka đơn giản, gồm:

Producer: phân loại message theo topic, sử dụng producer để publish message vào các topic. Dữ liệu được gửi đến partition của topic lưu trữ trên Broker. . *Broker*: Kafka cluster là một set các server, mỗi một set này được gọi là 1 broker *Consumer*: Kafka sử dụng consumer để subscribe vào topic, các consumer được định danh bằng các group name. Nhiều consumer có thể cùng đọc một topic.



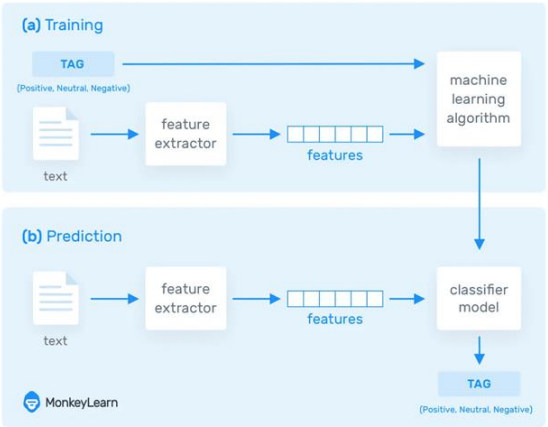
Chọn Confluent.io vì là một Kafka SaaS pay-as-you-go, chịu tải lớn, khả năng bảo mật và tích hợp. Thực hiện:

- Đăng ký và lấy các thông tin về Service (name, port), identity (username, password) ở confluent.io
- Tạo topic "Facebook_All_Comments"
- Tích hợp vào giai đoạn crawl bình luận, crawl được bình luận nào sẽ publish lên hệ thống tức thì bình luận ấy.

Xây dựng một service consumer listening ở dịch vụ Kafka trên với topic "Facebook_All_Comments". Khi nhận được một message, sẽ phân tích và lấy nội dung bình luận. Sau đó sẽ đưa bình luận này vào hệ thống phân tích NLP Tiếng Việt để xác định tích cực hay tiêu cực và ghi nhận vào Database SQL Server

4. Xây dựng Model phân loại cảm xúc

Phân loại văn bản (Text Classification) là bài toán thuộc nhóm học có giám sát (Supervised learning) trong học máy. Bài toán này yêu cầu dữ liệu cần có nhãn (label). Mô hình sẽ học từ dữ liệu có nhãn đó, sau đó được dùng để dự đoán nhãn cho các dữ liệu mới mà mô hình chưa gặp. Trong đề tài này, nhóm em tiến hành thu thập dữ liệu thô từ các bình luận trên Facebook. Sau đó dữ liệu thô được tiền xử lý và lấy mẫu, và gán nhãn trước khi tiến hành học máy. Dữ liệu lấy mẫu được chia thành ba nhóm: tập dữ liệu huấn luyện (training data), tập dữ liệu xác nhận (validation data) và tập dữ liệu kiểm tra (test data). Tập dữ liệu huấn luyện được sử dụng để thiết lập mô hình học máy.



Tổng quan mô hình

4.1 Chuẩn bị dữ liệu

Sử dụng các thư viện Beautiful Soup và Selenium trên ngôn ngữ Python để thu thập dữ liệu trên Facebook. Tập dữ liệu thu thập được có gồm các thông tin như tên cửa hàng, địa chỉ, tên khách hàng bình luận, thời gian bình luận, nội dung bình luận, tỷ lệ đánh giá của khách hàng đối với cửa hàng đó.

4.2 Tiền xử lý dữ liệu

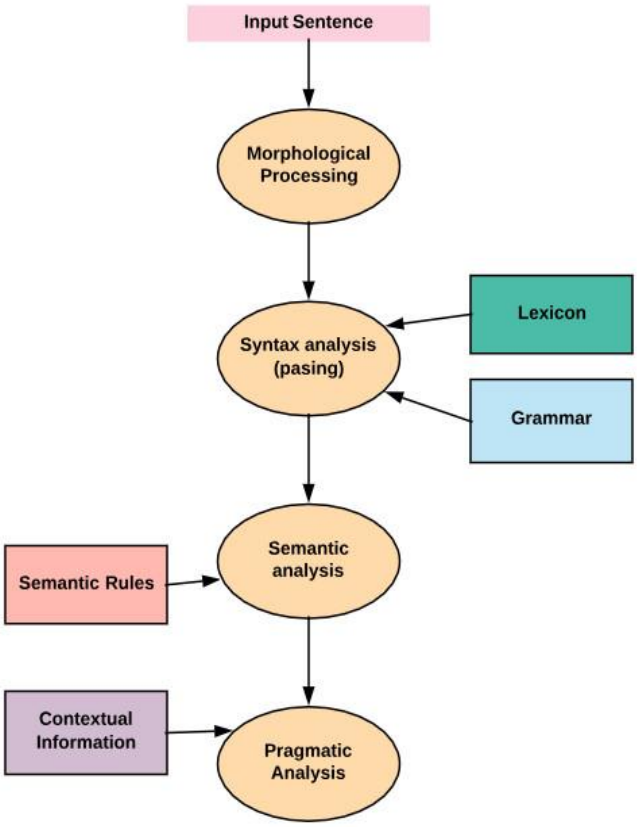
Dữ liệu thu thập về sẽ có dạng thô, do chưa qua xử lý nên có thể dữ liệu bị rỗng, dữ liệu sai chính tả, dữ liệu quá ngắn, quá dài hoặc chứa các biểu tượng icon. Điều này sẽ gây ảnh hưởng đến kết quả của việc phân tích, vì vậy cần làm sạch dữ liệu.

•*Xóa các icon, kí tự đặc biệt:* các kí tự đặc biệt không mang ý nghĩa phân loại, mặc khác sẽ gây nhiễu trong quá trình phân tích. Chuyển tất cả về chữ thường: mỗi số, ký tự đặc biệt, ký tự là đại diện cho một dãy nhị phân trong bộ nhớ máy tính. Chữ in hoa sẽ có mã Unicode khác chữ in thường, về mặt ngữ nghĩa là giống nhau tuy nhiên máy tính sẽ không thể phân biệt dữ liệu đầu vào, dẫn đến có thể kết quả dự đoán bị ảnh

hưởng. Vì vậy việc chuyển toàn bộ chữ về chữ thường là hợp lý cho hệ thống phân tích và dự đoán.

• *Chuyển dạng từ rõ nghĩa*: việc chuyển dạng từ rõ nghĩa là cần thiết cho bước tiền xử lý dữ liệu. Các bình luận trên Facebook do người dùng bình luận tiếng Việt nên việc viết tắt hoặc sai chính tả.… hay dữ liệu không đồng bộ, không chuẩn hóa. Việc này sẽ ảnh hưởng gây nhiều kết quả phân tích.

•*Xóa dòng dữ liệu:* tập dữ liệu thu về sẽ có nhiều dữ liệu bị trống, dữ liệu trống không có ý nghĩa trong quá trình phân tích, gây tốn bộ nhớ lưu trữ



Components of NLP

Sau khi thực hiện tiền xử lý dữ liệu xong, tiến hành xây dựng mô hình học máy cho trên dữ liệu đã tiền xử lý.

Xây dựng tập train/test

Dữ liệu dùng sau khi tiền xử lý được lưu thành 1 file duy nhất.

Sử dụng thư viện sklearn trong Python giúp tách dữ liệu làm 2 tập train/test riêng biệt và thực hiện các công việc sau:

- Đọc dữ liệu từ file và tách làm 2 list text (dữ liệu) và label (nhãn). Dữ liệu text[i] sẽ có nhãn là label[i].
- Chia làm 2 tập train (X_train, y_train) và test (X_test, y_test) theo tỉ lệ 80% train, 20% test.
- Lưu train/test data ra file để sử dụng cho việc train với thư viện Fasttext.
- Đưa label về dạng vector để tiện cho tính toán sử dụng LabelEncoder.

Với các thuật toán machine learning truyền thống sử dụng thư viện sklearn, xây dựng bộ trích xuất đặc trưng (feature extractor) sử dụng TF-IDF.

Phân loại văn bản với Naive Bayes

Phân loại văn bản với Logistic Regression

Phân loại văn bản với SVM

Phân loại văn bản dùng Fasttext

So sánh các mô hình cho kết quả tốt nhất trên tập test:

• Mô hình Logistic Regression và SVM cho kết quả vượt trội hơn Naive Bayes. Điều này do khả năng học của 2 mô hình này tốt hơn một mô hình ngây thơ (naive).

• Một số nhãn cho độ chính xác phân loại thấp (giải trí, nhịp sống, sống trẻ). Nguyên nhân có thể do các chuyên mục này không thực sự quá rõ ràng, nổi bật so với các chuyên mục khác nên mô hình dễ bị sai hơn.

• Mô hình sử dụng thư viện Fasttext cho kết quả thấp nhất (không như kỳ vọng của mình).

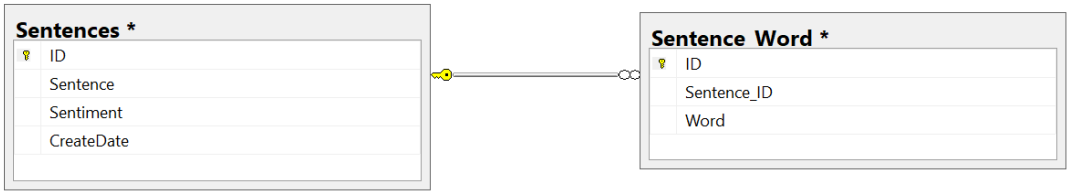
5. Cơ sở dữ liệu

5.1 Mô tả bảng

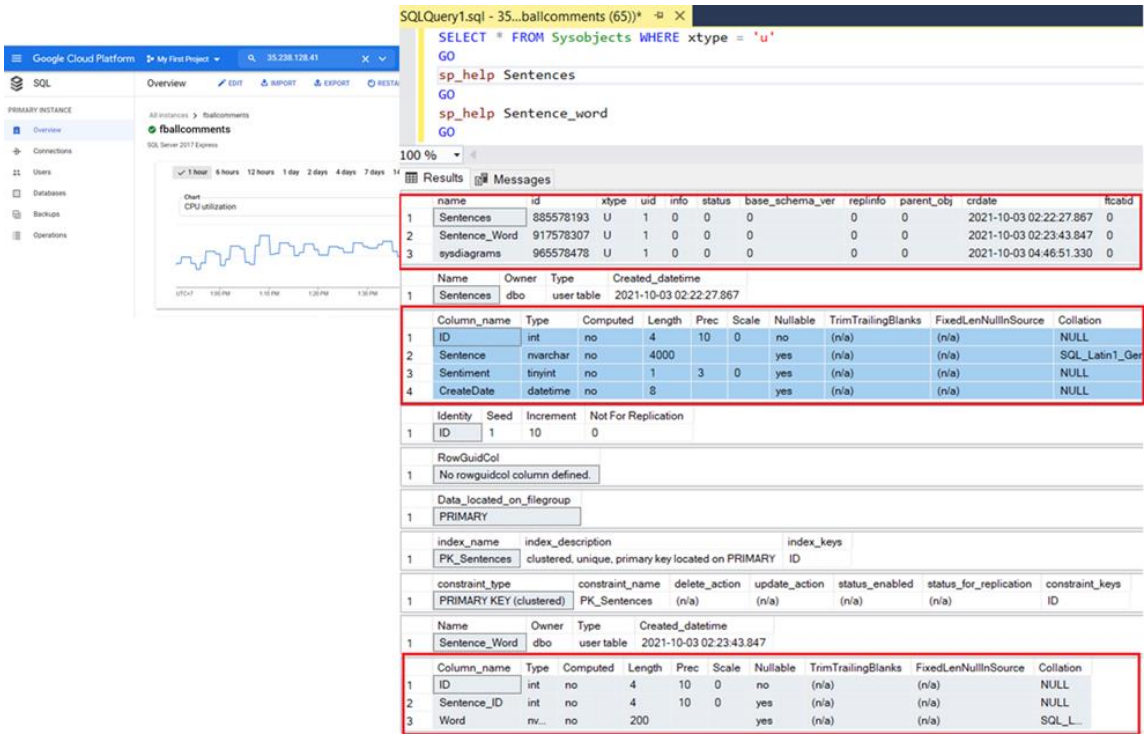
Cấu trúc cơ sở dữ liệu gồm 2 bảng chính

+ Sentences (ID , Sentence , Sentiment , CreateDate)
+ Sentences_word (id , sentence_ID , Word)

Lược đồ cơ sở dữ liệu

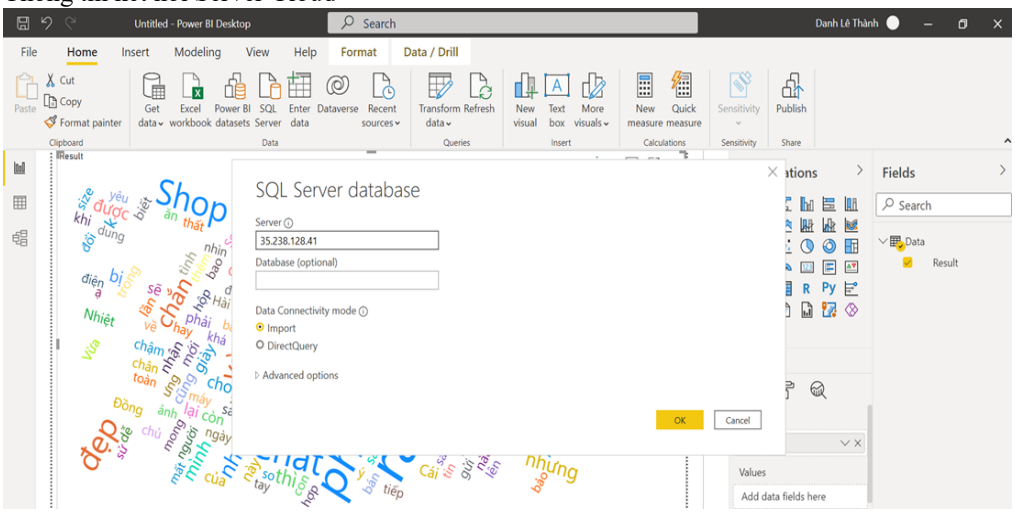


Mô tả trường trong bảng cơ sở dữ liệu

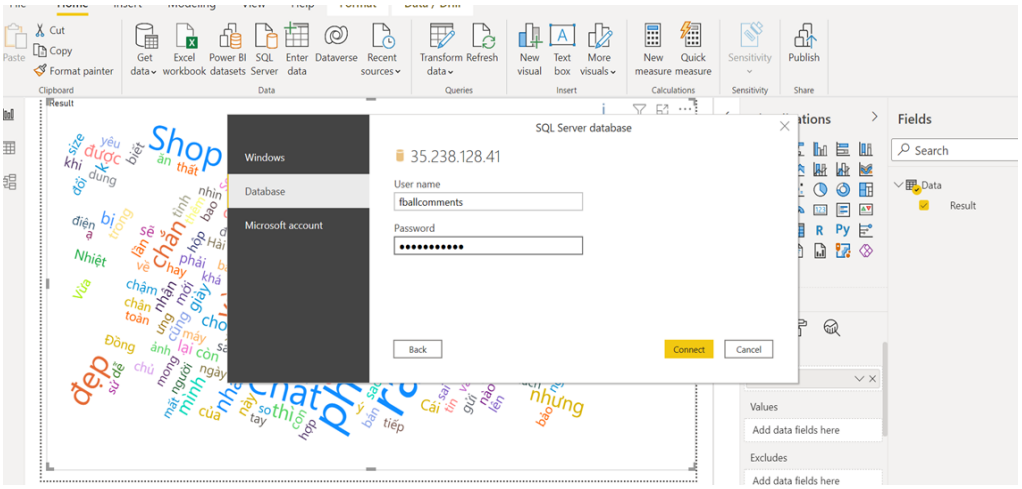


5.2 Kết nối SQL Server Cloud

Thông tin kết nối Server Cloud

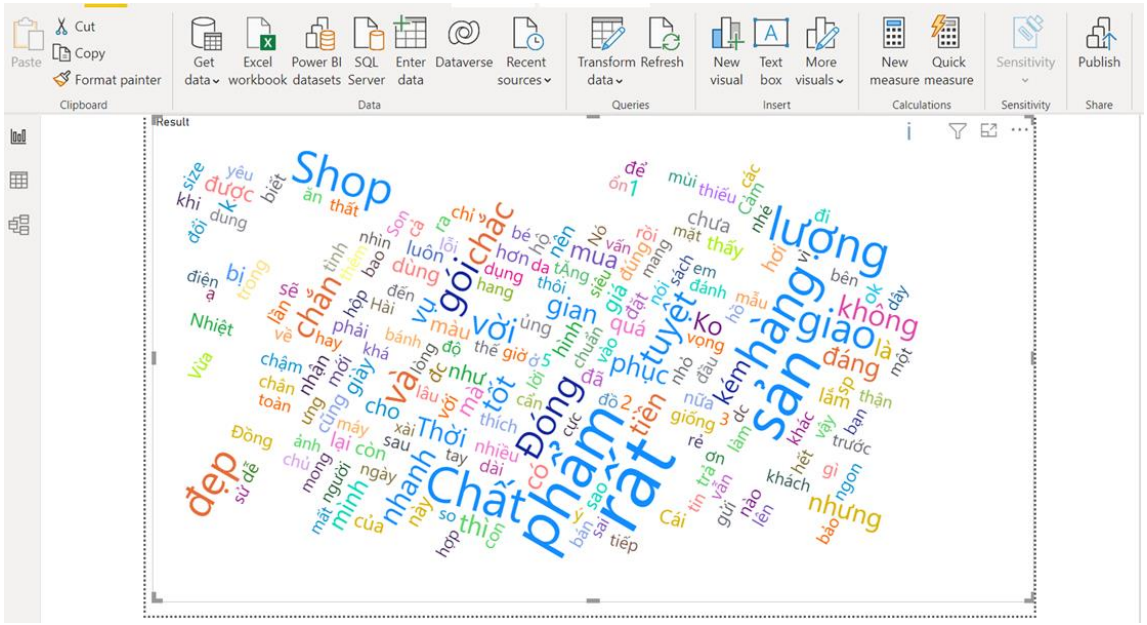


Nhập thông tin UserName , Password SQL Server Cloud



6. Power BI

Power BI là ứng dụng và dịch vụ trên nền tảng đám mây, có giao diện vô cùng thân thiện với người dùng. Nó là một công cụ BI và trực quan hóa dữ liệu từ các nguồn khác nhau để trở thành một bảng dashboard có thể tương tác trực tiếp và các báo cáo phân tích.
Power BI có thể kết nối với nhiều nguồn dữ liệu khác nhau, từ Excel cho đến các cơ sở dữ liệu trên đám mây, cũng như trên các ứng dụng.
Power BI trên màn hình desktop với (Power BI desktop), dịch vụ SaaS online tên là Power BI service và các ứng dụng Power BI trên hệ điều hành Windows, iOS và Android.
Sử dụng Power BI kết nối SQL Server Cloud và Visualize các từ xuất hiện theo tần suất từ nhiều đến thấp



7. Kết luận và hướng phát triển

Trong đồ án này, chúng em đã hoàn thành một giải pháp ứng dụng trong phân tích ngôn ngữ tự nhiên bằng tiếng việt, cụ thể là phân tích cảm xúc khách hàng dựa trên bình luận được đăng tải trên Facebook. Giải quyết được bài toán trong thời kỳ bùng nổ dữ liệu đó là cung cấp các thông tin trải nghiệm khách hàng sau khi mua hoặc sử dụng dịch vụ. Từ đó các thông tin đó doanh nghiệp sẽ có chiến lược để phát triển dịch vụ sản phẩm tốt hơn nhằm thu hút và giữ chân khách hàng. Ngoài ra đồ án này sẽ là tiền đề cho các ứng dụng phân tích dữ liệu, sử dụng giải pháp này để tích hợp vào các ứng dụng với mục đích khảo sát cảm xúc trải nghiệm khách hàng đối với tất cả sản phẩm dịch vụ khác nhau.

Trong thời gian tới nhóm chúng em sẽ cung cấp thông tin cụ thể hơn và dễ theo dõi hơn cho doanh nghiệp, thông qua việc ứng dụng triển khai các báo cáo phân tích ý kiến người dùng trên website, và đặc biệt là trên thiết bị di động, giúp doanh nghiệp tiện lợi hơn trong việc xem báo cáo và ra quyết định tốt hơn.

Tài liệu tham khảo

1. https://galaxyz.net/cach-thuc-hien-phan-tich-cam-xuc-trong-python-3-bang-bo cong-cu-ngon-ngu-tu-nhien-nltk.341.anews
2. https://ichi.pro/vi/lap-mo-hinh-chu-de-va-phan-tich-tinh-cam-tren-du-lieu-twitter-bang-spark-11082558915635
3. https://fraud-detection-handbook.github.io/fraud-detection handbook/Chapter_3_GettingStarted/BaselineModeling.html
4. https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_3_GettingStarted/SimulatedDataset.html