**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**NGUYỄN THÀNH THÁI**

**TRẦN ANH QUÂN**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG HỆ THỐNG PHÂN TÍCH TRẠNG THÁI CẢM XÚC CỦA THÀNH VIÊN TRÊN MẠNG XÃ HỘI**

**KỸ SƯ NGÀNH KỸ THUẬT PHẦN MỀM**

**TP. HỒ CHÍ MINH, 2015**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

**NGUYỄN THÀNH THÁI – 111520357**

**TRẦN ANH QUÂN – 11520305**

**KHÓA LUẬN TỐT NGHIỆP**

**XÂY DỰNG HỆ THỐNG PHÂN TÍCH TRẠNG THÁI CẢM XÚC CỦA THÀNH VIÊN TRÊN MẠNG XÃ HỘI**

**KỸ SƯ NGÀNH KỸ THUẬT PHẦN MỀM**

**GIẢNG VIÊN HƯỚNG DẪN**

**THS. NGUYỄN TRÁC THỨC**

**TP. HỒ CHÍ MINH, 2015**

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …………………… ngày ………………….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

* 1. …………………………………………. – Chủ tịch.
  2. …………………………………………. – Thư ký.
  3. …………………………………………. – Ủy viên.
  4. …………………………………………. – Ủy viên.

|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |
|  | *TP. HCM, ngày…..tháng…..năm……..* |

**NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP**

**(CỦA CÁN BỘ HƯỚNG DẪN)**

**Tên khóa luận:**

**XÂY DỰNG HỆ THỐNG PHÂN TÍCH TRẠNG THÁI CẢM XÚC CỦA THÀNH VIÊN TRÊN MẠNG XÃ HỘI**

**Nhóm SV thực hiện: Cán bộ hướng dẫn:**

Nguyễn Thành Thái 11520357 ThS. Nguyễn Trác Thức

Trần Anh Quân 11520305

**Đánh giá Khóa luận**

1. Về cuốn báo cáo:

Số trang Số chương

Số bảng số liệu Số hình vẽ

Số tài liệu tham khảo Sản phẩm

Một số nhận xét về hình thức cuốn báo cáo:

1. Về nội dung nghiên cứu:

1. Về chương trình ứng dụng:

1. Về thái độ làm việc của sinh viên:

**Đánh giá chung:**

**Điểm từng sinh viên:**

Nguyễn Thành Thái **:………../10**

Trần Anh Quân **:………../10**

**Người nhận xét**

(Ký tên và ghi rõ họ tên)

|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |
|  | *TP. HCM, ngày…..tháng…..năm……..* |

**NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP**

**(CỦA CÁN BỘ PHẢN BIỆN)**



|  |  |  |
| --- | --- | --- |
| **Tên khóa luận:** | | |
| **XÂY DỰNG HỆ THỐNG PHÂN TÍCH TRẠNG THÁI CẢM XÚC CỦA THÀNH VIÊN TRÊN MẠNG XÃ HỘI** | | |
| **Nhóm SV thực hiện:** | | **Cán bộ phản biện:** |
| Nguyễn Thành Thái 11520357 |  |  |
| Trần Anh Quân 11520305 |  |  |

**Đánh giá Khóa luận**

1. Về cuốn báo cáo:

Số trang Số chương

Số bảng số liệu Số hình vẽ

Số tài liệu tham khảo Sản phẩm

Một số nhận xét về hình thức cuốn báo cáo:

1. Về nội dung nghiên cứu:

1. Về chương trình ứng dụng:

1. Về thái độ làm việc của sinh viên:

**Đánh giá chung:**

**Điểm từng sinh viên:**

Nguyễn Thành Thái **:………../10**

Trần Anh Quân **:………../10**

**Người nhận xét**

(Ký tên và ghi rõ họ tên)

|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |



**ĐỀ CƯƠNG CHI TIẾT**

|  |  |
| --- | --- |
| **Tên đề tài: XÂY DỰNG HỆ THỐNG PHÂN TÍCH TRẠNG THÁI CẢM XÚC CỦA THÀNH VIÊN TRÊN MẠNG XÃ HỘI** | |
| **Cán bộ hướng dẫn: ThS. Nguyễn Trác Thức** | |
| **Thời gian thực hiện:** Từ ngày 15/09/2015 đến ngày 15/01/2015 | |
| **Sinh viên thực hiện:**  **Nguyễn Thành Thái – 11520357**  **Trần Anh Quân – 11520305** | |
| **Nội dung đề tài:**Xây dựng hệ thống phân tích trạng thái cảm xúc của thành viên trên mạng xã hội. | |
| **Kế hoạch thực hiện:**  *Khóa luận được thực hiện trong 16 tuần, nhóm chúng em chia công việc theo tuần để dễ dàng đánh giá kết quả. Kế hoạch thực hiện cụ thể như sau:*   |  |  |  | | --- | --- | --- | | **Tuần** | **Công việc** | **Người thực hiện** | | 1-2(2 tuần) | * Tìm hiểu thuật toán LDA (Latent Dirichlet Allocation) * Tìm hiểu các thuật toán và giải pháp đánh giá cảm xúc Tiếng Việt đã có. * Khảo sát và xác định ứng dụng cụ thể sẽ làm. | Thái + Quân | | 3(1 tuần) | * Phát thảo đề cương các nội dung tìm hiểu * Mô tả chi tiết ý tưởng ứng dụng và mô tả các dữ liệu sử dụng | Thái + Quân | | 4-5(2 tuần) | * Chuẩn bị dữ liệu phù hợp với LDA Model của Spark * Cài đặt thử nghiệm thuật toán LDA * Thiết kế kiến trúc hệ thống | Thái | | * Tìm hiểu API RestFB để lấy thông tin từ các trang Facebook * Thiết kế dữ liệu, thiết kế database để lưu trữ dữ liệu lấy được thì Facebook | Quân | | 6-7(2 tuần) | * Xử lý dữ liệu thô ban đầu từ Facebook thành các tài liệu được chuẩn hóa | Quân | | * Xử lý các biểu tượng cảm xúc thành các từ ngữ cảm xúc có nghĩa cho bước xử lý tiếp theo | Thái | | 8-10(2 tuần) | * Thiết kế giao diện hệ thống để nhập các lựa chọn và hiển thị thông tin theo yêu cầu ban đầu | Thái | | * Sử dụng và cải tiến bộ từ điển cảm xúc VietSentiWordNet cho đánh giá cảm xúc trên mạng xã hội | Quân | | 11(1 tuần) | * Cài đặt hệ thống sử dụng các thành phần đã chuẩn bị và dựa trên thiết kế đã có. * Xây dựng vector đặc trưng dựa vào TF-IDF, tạo dữ liệu cho SVM, tiến hành phân lớp cảm xúc. | Thái | | * Sử dụng bộ công cụ phân lớp SVM để phân lớp các dữ liệu đã chuẩn hóa dựa trên kết quả đầu ra của việc đánh giá cảm xúc bằng VietSentiWordNet | Quân | | 12(1 tuần) | * Kiểm lỗi hệ thống và sửa lỗi nếu có | Thái + Quân | | 13–14(2 tuần) | * Thời gian dự trữ cho việc viết báo cáo và sửa lỗi nâng cấp ứng dụng | Thái + Quân | | 15–16(2 tuần) | * Hoàn tất báo cáo. | Thái + Quân | | |
| **Xác nhận của CBHD**  (Ký tên và ghi rõ họ tên) | **TP. HCM, ngày….tháng …..năm…..**  **Sinh viên**  (Ký tên và ghi rõ họ tên) |

**LỜI CÁM ƠN**

Để có thể hoàn thành khóa luận một cách hoàn chỉnh, nhóm đã nhận được rất nhiều sự hướng dẫn nhiệt tình của quý Thầy Cô, cũng như sự động viên, ủng hộ của gia đình và bạn bè trong suốt thời gian thực hiện. Vì vậy, nhóm chúng em xin phép được gửi những lời cảm ơn chân thành nhất.

Đầu tiên, nhóm xin chân thành cảm ơn và bày tỏ lòng biết ơn đến thầy Nguyễn Trác Thức, thầy đã tận tình chỉ bảo, hướng dẫn và động viên nhóm chúng em để nhóm có thể hoàn thành khóa luận một cách tốt đẹp nhất. Nhóm chúng em xin gởi đến thầy những lời tri ân nhất đối với những điều mà thầy đã dành cho nhóm.

Nhóm cũng xin chân thành cảm ơn toàn thể quý thầy cô trong khoa Công Nghệ Phần Mềm và toàn thể quý thầy cô của Trường Đại Học Công Nghệ Thông Tin TP.HCM. Quý thầy cô đã luôn tận tình truyền đạt những kiến thức, kinh nghiệm quý báu cũng như tạo mọi điều kiện thuận lợi nhất cho chúng em trong suốt quá trình học tập tại trường và đặc biệt hơn là trong suốt quá trình thực hiện khóa luận này.

Xin chân thành bày tỏ lòng biết ơn sâu sắc đến gia đình, bạn bè những người đã không ngừng động viên, hỗ trợ và tạo mọi điều kiện tốt nhất cho chúng em trong suốt thời gian thực hiện khóa luận.

Sau cùng, nhóm xin kính chúc toàn thể quý thầy cô trong khoa Công Nghệ Phần Mềm và thầy Nguyễn Trác Thức luôn dồi dào sức khỏe, sự nhiệt huyết và thành công trong cuộc sống cũng như trong sứ mệnh truyền đạt kiến thức cho các thế hệ trẻ mai sau.

Trân trọng

Nhóm sinh viên thực hiện

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU 4](#_Toc440482216)

[1.1 Giới thiệu đề tài 4](#_Toc440482217)

[1.2 Mục tiêu đề tài 5](#_Toc440482218)

[1.3 Đối tượng nghiên cứu 5](#_Toc440482219)

[1.4 Phạm vi nghiên cứu 5](#_Toc440482220)

[1.5 Nội dung 6](#_Toc440482221)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 7](#_Toc440482222)

[2.1 Mô hình chủ đề ẩn (LDA) 7](#_Toc440482223)

[2.1.1 Giới thiệu 7](#_Toc440482224)

[2.1.2 Đặc điểm của LDA 8](#_Toc440482225)

[2.1.3 Mô hình sinh trong LDA 8](#_Toc440482226)

[2.1.4 Ước lượng tham số và suy luận 10](#_Toc440482227)

[2.2 Phân tích và đánh giá cảm xúc 13](#_Toc440482232)

[2.2.1 Định nghĩa quan điểm 13](#_Toc440482233)

[2.2.2 Bài toán phân tích cảm xúc 14](#_Toc440482234)

[2.2.3 Các nghiên cứu về và công cụ cho việc phân tích cảm xúc đã có 16](#_Toc440482235)

[2.2.4 Các phương pháp tiếp cận cho việc phân tích cảm xúc 19](#_Toc440482236)

[2.2.5 Đánh giá cảm xúc dựa vào từ vựng (Lexicon) và vấn đề gặp phải 22](#_Toc440482237)

[2.2.6 Đề xuất về mô hình phân tích cảm xúc kết hợp phương pháp phân lớp Lexicon và phương pháp SVM. 23](#_Toc440482238)

[2.3 Giới thiệu thư viện hỗ trợ lấy dữ liệu từ Facebook - RestFB 25](#_Toc440482239)

[2.4 Giới thiệu công cụ tách từ Tiếng Việt – vnTokenizer 27](#_Toc440482240)

[2.5 Giới thiệu về kho ngữ liệu WordNet 28](#_Toc440482242)

[2.5.1 Giới thiệu về kho ngữ liệu SentiWordNet và VietSentiWordNet 29](#_Toc440482243)

[2.5.2 Một số khái niệm được sử dụng về WordNet 29](#_Toc440482244)

[2.5.3 Các phiên bản SentiWordNet 31](#_Toc440482245)

[2.5.4 Giới thiệu bộ từ điển biểu tượng cảm xúc (emoticon) 31](#_Toc440482246)

[2.6 Xây dựng vector đặc trưng (TF-IDF): 32](#_Toc440482247)

[2.7 Giới thiệu về phương pháp phân lớp SVM 33](#_Toc440482252)

[2.7.1 Ý tưởng của phương pháp 33](#_Toc440482253)

[2.7.2 Không gian đặc trưng 33](#_Toc440482254)

[2.7.3 Cơ sở lý thuyết của phương pháp SVM 34](#_Toc440482255)

[2.7.4 Các bước chính của phương pháp SVM 35](#_Toc440482256)

[2.8 Đánh giá độ chính xác của hệ thống 35](#_Toc440482257)

[2.9 Giới thiệu về Apache Spark 36](#_Toc440482259)

[2.9.1 Tổng quan 36](#_Toc440482260)

[2.9.2 Hadoop và Spark 37](#_Toc440482261)

[2.9.3 Những tính năng nổi bật của Spark 39](#_Toc440482262)

[2.9.4 Hệ sinh thái của Apache Spark 39](#_Toc440482263)

[CHƯƠNG 3. XÂY DỰNG HỆ THỐNG TÌM KIẾM CHỦ ĐỀ ẨN VÀ ĐÁNH GIÁ CẢM XÚC TRÊN TRANG FACEBOOK 42](#_Toc440482264)

[3.1 Quy trình phát triển phần mềm 42](#_Toc440482265)

[3.1.1 Mô hình phát triển phần mềm 42](#_Toc440482266)

[3.1.2 Xác định yêu cầu 43](#_Toc440482267)

[3.1.3 Kiến trúc hệ thống 44](#_Toc440482268)

[3.2 Phân tích yêu cầu 45](#_Toc440482285)

[3.2.1 Lược đồ use case 45](#_Toc440482286)

[3.2.2 Mô tả lược đồ use case 45](#_Toc440482287)

[3.2.3 Sơ đồ luồng dữ liệu toàn hệ thống 48](#_Toc440482288)

[3.2.4 Chức năng tìm chủ đề ẩn 50](#_Toc440482289)

[3.2.5 Chức năng phân tích cảm xúc 52](#_Toc440482290)

[3.3 Thiết kế màn hình 55](#_Toc440482291)

[3.3.1 Màn hình chính 55](#_Toc440482292)

[3.3.2 Màn hình lấy dữ liệu từ Facebook 56](#_Toc440482293)

[3.3.3 Màn hình chọn dữ liệu để phân tích 57](#_Toc440482294)

[3.3.4 Màn hình xem các chủ để đang trao đổi 58](#_Toc440482295)

[3.3.5 Màn hình xem thống kê về đánh giá cảm xúc 59](#_Toc440482296)

[3.4 Thực nghiệm và đánh giá 60](#_Toc440482298)

[3.4.1 Môi trường thực nghiệm 60](#_Toc440482299)

[3.4.2 Chuẩn bị dữ liệu 61](#_Toc440482300)

[3.4.3 Kết quả thực nghiệm 62](#_Toc440482301)

[CHƯƠNG 4. KẾT LUẬN 69](#_Toc440482303)

[4.1 Kết quả đạt được 69](#_Toc440482304)

[4.2 Hướng phát triển 70](#_Toc440482307)

DANH MỤC HÌNH VẼ

Hình 2.1: Giới thiệu về LDA 7

Hình 2.2: Mô hình sinh của LDA 9

Hình 2.3: Trang web Twitter Sentiment 18

Hình 2.4: Mô hình phân lớp cảm xúc đề xuất 23

Hình 2.5 : Mô tả Social Graph 25

Hình 2.6: Mô tả kết nối với cơ sở dữ liệu của Facebook 27

Hình 2.7: Cấu trúc trong file SentiWordNet3.0 30

Hình 2.8: Ánh xạ vào không gian đặc trưng 34

Hình 2.9: Ba cách để cài đặt Spark trên nền tảng Hadoop 39

Hình 2.10: Hệ sinh thái Spark 40

Hình 3.1: Mô hình thác nước cải tiến 43

Hình 3.2: Kiến trúc hệ thống 44

Hình 3.3: Lược đồ Use Case 45

Hình 3.4: Sơ đồ DFD cho toàn hệ thống 48

Hình 3.5: Sơ đồ DFD cho chức năng tìm chủ đề ần 50

Hình 3.6: Sơ đồ Sequence cho chức năng tìm chủ đề ẩn 51

Hình 3.7: Sơ đồ DFD cho chức năng phân lớp cảm xúc 52

Hình 3.8: Sơ đồ Sequence giai đoạn huấn luyện cho SVM 54

Hình 3.9: Sơ đồ Sequence giai đoạn phân lớp cảm xúc 54

Hình 3.10: Màn hình chính 55

Hình 3.11: Màn hình lấy dữ liệu từ Facebook 56

Hình 3.12: Màn hình chọn dữ liệu để phân tích 57

Hình 3.13: Màn hình xem các chủ đề đang được trao đổi 58

Hình 3.14: Màn hình xem thống kê đánh giá cảm xúc 59

Hình 3.15: Biểu đồ giá trị Log-likelihood theo số lượng chủ đề 63

Hình 3.16: Đánh giá kết quả phân lớp với TDL1 64

Hình 3.17: Đánh giá kết quả phân lớp với TDL2 65

Hình 3.18: Đánh giá kết quả phân lớp với TDL3 66

Hình 3.19: Đánh giá kết quả phân lớp với TDL4 67

DANH MỤC BẢNG

Bảng 2.1: So sánh các giải pháp đã có cho việc phân tích cảm xúc 20

Bảng 2.2: Mô tả cấu trúc của SentiWordNet 3.0 30

Bảng 2.3: Các kí hiệu sử dụng trong đánh giá độ chính xác 35

Bảng 3.1: Cấu hình phần cứng cho thực nghiệm 60

Bảng 3.2: Công cụ và thư viện hỗ trợ 60

**DANH MỤC TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| **Từ viết tắt** | **Ý nghĩa** |
| LDA | Latent Dirichlet Allocation |
| pLSI | Probabilistic Latent Semantic Indexing |
| TF | Term Frequencies |
| IDF | Inverse Document Frequency |
| SVM | Support Vector Machine |
| RMI | Remote Method Invocation |
| TDL | Tập dữ liệu |

TÓM TẮT KHÓA LUẬN

Cảm xúc là một hình thức trải nghiệm cơ bản của con người về thái độ của chính mình đối với sự vật, hiện tượng, đối với người khác hoặc với bản thân. Cảm xúc luôn là một phần không thể thiếu của cuộc sống, nó là cái nhìn, là cách đánh giá nhanh nhất của mỗi người với những điều mình quan tâm. Phân tích cảm xúc là một bài toán quan trọng của việc khai phá quan điểm. Hệ thống phân tích cảm xúc, phải là một hệ thống có khả năng tự động phân loại các đánh giá, các bình luận theo hướng tích cực, tiêu cực hay trung tính của quan điểm. Với sự phát triển và bùng nổ của mạng xã hội, hệ thống phân tích cảm xúc là cần thiết để giúp chúng ta phần nào nắm được thái độ, quan điểm của các thành viên mạng xã hội thông qua các thông tin bình luận và được gom nhóm lại thành các chủ đề khác nhau.

Có rất nhiều nghiên cứu trong và ngoài nước về phương pháp phân lớp quan điểm, chủ yếu là theo hai hướng chính: phương pháp học máy và phương pháp ngữ nghĩa dựa vào độ đo thông tin. Trong khuôn khổ của khóa luận, chúng tôi sẽ trình bày khả năng biểu diễn tập dữ liệu là các bài đăng (status) và các bình luận (comments) của các thành viên mạng xã hội Facebook thành các chủ đề được bàn tán nhiều nhất sử dụng mô hình chủ đề ẩn LDA (Latent Dirichlet allocation). Sau đó là thực nghiệm mô hình kết hợp khả năng phân tích cảm xúc dựa vào các từ điển cảm xúc đã có, kết hợp với phương pháp phân lớp SVM (Support vector machine) để tự động phân lớp các bình luận vào lớp cảm xúc tương ứng. Các kết quả đánh giá và phân lớp dữ liệu được biểu diễn dưới dạng biểu đồ, thực nghiệm trên miền dữ liệu từ các fanpage của Facebook với gần 5000 bình luận, cho kết quả độ chính xác cao nhất là 79%.

Từ khóa: sentiment analysis, social network, bigdata, topic modeling, machine learning

MỞ ĐẦU

Mạng xã hội ngày nay được giới trẻ rất ưa dùng, vậy “**Mạng xã hội**” là gì? Theo wikipedia Mạng xã hội, hay còn gọi là mạng xã hội ảo (social network) là dịch vụ nối kết các thành viên cùng sở thích trên Internet lại với nhau với nhiều mục đích khác nhau, không phân biệt không gian và thời gian. Những người tham gia vào mạng xã hội còn được gọi là cư dân mạng.

Mạng xã hội cung cấp những tính năng như chat, e-mail, phim ảnh, voice chat, chia sẻ file, viết blog và xã luận. Mạng xã hội đang từng bước thay đổi hoàn toàn cách cư dân mạng liên kết với nhau và trở thành một phần tất yếu của mỗi ngày cho hàng trăm triệu thành viên khắp thế giới. Cách dịch vụ mạng xã hội này có nhiều phương cách để các thành viên tìm kiếm bạn bè, đối tác: dựa theo group, dựa theo thông tin cá nhân, dựa trên sở thích (như thể thao, phim ảnh, sách báo hoặc ca nhạc), cũng như theo các lĩnh vực quan tâm: kinh doanh, mua bán...

Hiện nay, thế giới có hàng trăm mạng xã hội khác nhau như Facebook, Twitter, Google Plus, Linkedin, MySpace,... hay tại Việt Nam có Zalo, ZingMe, Yume,... đã và đang kết nối hàng tỉ người theo từng phút, từng giây thông qua đường truyền internet.

Facebook với một lượng người dùng cực kì lớn, và phổ biến nhất tại Việt Nam, sẽ là nguồn dữ liệu cực kì phong phú và hữu ích cho nhóm thực hiện đề tài nghiên cứu “Phân tích cảm xúc của các thành viên trên mạng xã hội”.

# GIỚI THIỆU

## Giới thiệu đề tài

Khóa luận sẽ xây dựng một hệ thống thực hiện việc tổng hợp các chủ đề đang được trao đổi trên mạng xã hội Facebook cụ thể là từ các trang công khai trên Facebook và đưa ra các đánh giá cảm xúc của các thành viên đối với từng chủ đề theo ba trạng thái cảm xúc: Vui vẻ - tích cực, trạng thái trung tính và trạng thái buồn chán – tiêu cực.

Phân tích tâm lý hoặc đánh giá cảm xúc đề cập đến ứng dụng xử lý ngôn ngữ tự nhiên, ngôn ngữ học tính toán, và phân tích văn bản để xác định và trích xuất thông tin chủ quan từ các tài liệu. Nói chung, phân tích tâm lý nhằm mục đích để xác định thái độ của một người hoặc một nhóm người đối với một số chủ đề hoặc phân cực tổng thể (các ý kiến trái chiều) theo ngữ cảnh của một tài liệu. Thái độ có thể là ý kiến hoặc đánh giá (theo thuyết thẩm định), trạng thái tình cảm (là ý kiến, hiệu ứng cảm xúc của tác giả khi viết), hoặc các thông tin cảm xúc mong đợi (là ý kiến, hiệu ứng cảm xúc của tác giả mong muốn có ở người đọc).

Đánh giá cảm xúc của các thành viên trên mạng xã hội (đánh giá tình cảm, trích rút quan điểm / tình cảm) là lĩnh vực nghiên cứu nhằm cố gắng làm cho hệ thống tự động xác định quan điểm mà người dùng bày tỏ thông qua các bình luận và các ý kiến cá nhân từ văn bản được viết bằng ngôn ngữ tự nhiên.

***Tại sao phải thực hiện phân tích và đánh giá cảm xúc?***

* Các quan điểm có ảnh hưởng vô cùng lớn đến các hành vi của chúng ta.
* Niềm tin và nhận thức của chúng ta trong thực tế là phụ thuộc vào cách thức những người khác nhìn thế giới.
* Bất cứ khi nào chúng ta phải đưa ra một quyết định, chúng ta thường phải tham khảo những ý kiến của những người khác.
  + Với cá nhân: Tham khảo ý kiến từ bạn bè, gia đình.
  + Với tổ chức: Sử dụng các cuộc khảo sát, điều tra, các cuộc tham dò dư luận, tư vấn.

## Mục tiêu đề tài

Tìm hiểu các thuật toán về phân cụm tài liệu theo chủ đề cũng như các thuật toán và giải pháp để phân lớp cảm xúc đã có. Trên cơ sở đó, đề xuất một mô hình để kết hợp các thuật toán và ứng dụng vào việc phân cụm các chủ đề, sau đó tiến hành đánh giá cảm xúc của những bài đăng/bình luận trên mạng xã hội Facebook theo từng chủ đề đã được tổng hợp.

Sản phẩm của khóa luận sẽ giúp cho những cơ quan quản lý, phòng ban hoặc doanh nghiệp cơ bản nắm được mức độ quan tâm của cộng đồng trên mạng xã hội về các chủ đề cũng như những phản ứng tích cực/tiêu cực về các chủ đề đó và đưa ra những điều chỉnh và phản ứng kịp thời.

## Đối tượng nghiên cứu

* Mạng xã hội Facebook.
* Thuật toán LDA dùng để tìm chủ đề ẩn và phân cụm chủ đề.
* Mô hình SentiWordNet và thư viện VietSentiWordNet.
* Xây dựng vector đặc trưng dựa trên kỹ thuật thống kê TF-IDF.
* Phương pháp phân lớp tài liệu SVM.

## Phạm vi nghiên cứu

Khóa luận đi sâu vào nghiên cứu kĩ thuật xây dựng module crawler để lấy dữ liệu từ mạng xã hội Facebook, nghiên cứu thuật toán TFIDF để vector hóa dữ liệu là các dòng trạng thái (status) và các bình luận (comment) của các thành viên trên mạng xã hội Facebook thành các vector đặc trưng. Từ dữ liệu là các vector đăng trưng này, khóa luận nghiên cứu thuật toán tìm chủ đề ẩn (LDA) sử dụng dữ liệu đầu vào là các vector đặc trưng để tìm các chủ đề đang được trao đổi ở các trang cộng đồng trên mạng xã hội Facebook. Sau đó, khóa luận sử dụng thư viện VietSentiWordNet [1] cho đánh giá cảm xúc kết hợp với phương pháp phân lớp tài liệu SVM để tiến hành phân loại cảm xúc người dùng theo các chủ đề đang được trao đổi.

## Nội dung

Báo cáo gồm có 4 chương:

**Chương 1:**Giới thiệu mục tiêu, phạm vi, đối tượng nghiên cứu của đề tài.

**Chương 2:**Giới thiệu phương pháp tìm chủ đề ẩn sử dụng thuật toán LDA, giới thiệu và trình bày những nội dung cơ bản về phân phối Dirichlet ẩn và lấy mẫu Gibbs. Trình bày về mô hình SentiWordNet và thư viện VietSentiWordNet để đánh giá cảm xúc kết hợp với phương pháp phân cụm SVM. Đề xuất mô hình tìm chủ đề ẩn và mô hình đánh giá cảm xúc kết hợp thư viện VietSentiWordNet và SVM. Giới thiệu một số thư viện và công cụ hỗ trợ mà khóa luận sử dụng.

**Chương 3:** Thiết kế, xây dựng ứng dụng web mô phỏng chức năng tìm các chủ đề mà một trang cộng đồng trên Facebook đang trao đổi trong một khoảng thời gian nhất định và đánh giá cảm xúc về các chủ đề đó. Thực nghiệm và đánh giá kết quả.

**Chương 4:** Kết luận, trình bày những điểm mà khóa luận đạt được và hướng phát triển.

**Kết luận chương một:**

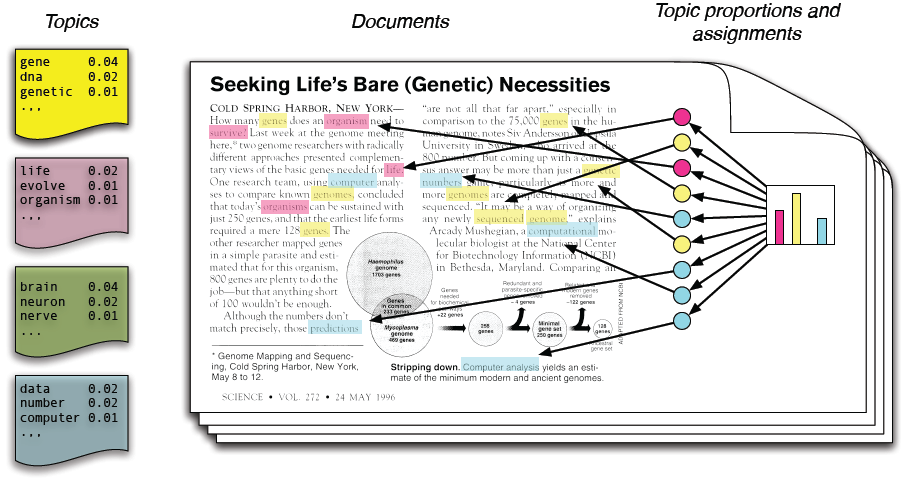
Trong chương này, khóa luận giới thiệu một cách tổng quan về những mục tiêu, nhiệm vụ, cơ sở và công cụ để nhóm thực hiện đề tài: Xây dựng hệ thống phân tích cảm xúc của thành viên trên Mạng xã hội. Khóa luận cũng nêu rõ phạm vi nghiên cứu của đề tài, tập trung vào những yêu cầu cần giải quyết để có thể đánh giá kết quả thực hiện cuối cùng.

Chương tiếp theo, khóa luận sẽ đi sâu vào phân tích bài toán khai phá cảm xúc, các khái niệm về cảm xúc, các nghiên cứu đã có cũng nhữ kỹ thuật và phương pháp để hiện thực mục tiêu của đề tài. Qua những tìm hiểu đó, khóa luận cũng sẽ đề xuất mô hình cụ thể để có thể áp dụng vào bài toán phân tích cảm xúc.

# CƠ SỞ LÝ THUYẾT

## Mô hình chủ đề ẩn (LDA)

### Giới thiệu

Hiện nay có nhiều nghiên cứu để khai thác một cách hiệu quả mối quan hệ giữa các dữ liệu với nhau. Trong đó, mô hình chủ đề ẩn là một bước tiến quan trọng trong việc mô hình hóa dữ liệu văn bản. Chúng được xây dựng dựa trên ý tưởng rằng mỗi tài liệu có một xác suất phân phối vào các chủ đề, và mỗi chủ đề là sự phân phối kết hợp giữa các từ khóa với nhau. Biểu diễn các từ và tài liệu dưới dạng phân phối xác suất có lợi ích rất lớn so với không gian vector thông thường.

Hình 2.1: Giới thiệu về LDA

Ý tưởng của các mô hình chủ đề ẩn là xây dựng những tài liệu mới dựa theo phân phối xác suất. Trước hết, để tạo ra một tài liệu mới, cần chọn ra một phân phối những chủ đề cho tài liệu đó, điều này có nghĩa tài liệu được tạo nên từ những chủ đề khác nhau, với những phân phối khác nhau. Tiếp đó, để sinh các từ cho tài liệu ta có thể lựa chọn ngẫu nhiên các từ vào phân phối xác suất của các từ trên các chủ đề.

Một cách hoàn toàn ngược lại, cho một tập tài liệu, có thể xác định một tập các chủ đề ẩn cho mỗi tài liệu và phân phối xác suất của các từ trên từng chủ đề. Có nhiều phương pháp về phân tích chủ đề sử dụng mô hình ẩn là Probabilistic Latent Semantic Indexing (pLSI) và Latent Dirichlet Allocation (LDA).

pLSI là một kỹ thuật thống kê nhằm phân tích những dữ liệu xuất hiện đồng thời. Phương pháp này được phát triển dựa trên LSA, mặc dù pLSI là một bước quan trọng trong việc mô hình hóa dữ liệu văn bản, tuy nhiên nó vẫn còn chưa hoàn thiện ở chỗ chưa xây dựng được một mô hình xác suất tốt ở mức độ tài liệu. Điều đó dẫn đến vấn đề gặp phải khi tìm phân phối xác suất cho một tài liệu nằm ngoài tập dữ liệu học, ngoài ra số lượng các tham số có thể tăng lên một cách tuyến tính khi kích thước của tập dữ liệu tăng.

### Đặc điểm của LDA

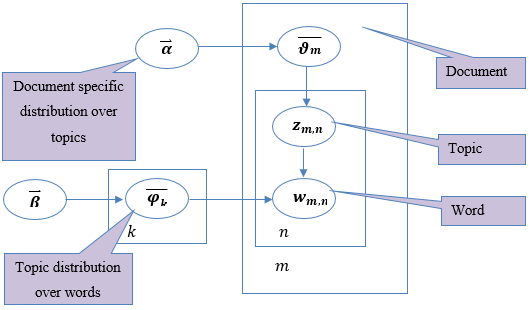
LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc. LDA được xây dựng dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủ đề (topic). Về bản chất, LDA là một mô hình Bayesian 3 cấp (three-level hierarchical Bayes model: corpus level, document level, word level). Mỗi tài liệu trong tập hợp được coi là một hỗn hợp xác định trên tập cơ bản các chủ đề. Về khía cạnh mô hình hóa văn bản, các xác suất chủ đề là một biểu diễn cụ thể, rõ ràng cho một tài liệu.

LDA là một mô hình hoàn thiện hơn so với pLSI và có thể khắc phục được những nhược điểm ở trên. Mô hình chủ đề ẩn này sẽ được sử dụng trong việc xây dựng hệ thống tìm ra các chủ đề từ mạng xã hội của khóa luận.

### Mô hình sinh trong LDA

Cho trước tập M tài liệu **D = {d1, d2... dM},** trong đó tài liệu thứ ***m*** gồm ***Nm*** từ, các từ ***wi*** được rút ra từ tập các thuật ngữ **{t1, t2... tV}**, **V** là số lượng của thuật ngữ. Mô hình LDA được biểu diễn dưới dạng đồ họa trong hình 2.2.

Quá trình sinh trong mô hình LDA diễn ra như sau: Mô hình LDA sinh các từ ***wm,n*** có thể quan sát, các từ này được phân chia về các tài liệu. Với mỗi tài liệu, một tỉ lệ chủ đề được chọn từ phân phối Dirichlet (*Dir* ()), từ đó, xác định các từ thuộc chủ đề cụ thể. Sau đó, với mỗi từ thuộc tài liệu, chủ đề của từ đó được xác định là một chủ đề cụ thể bằng cách lấy mẫu từ phân bố đa thức (*Mult* ()). Cuối cùng, từ phân bố đa thức (*Mult* ()), một từ cụ thể ***wm,n*** được sinh ra dựa trên chủ đề đã được xác định. Các chủ đề được lấy mẫu một lần trong toàn kho ngữ liệu.



Hình 2.2: Mô hình sinh của LDA

Các khối vuông trong hình trên biểu diễn các quá trình lặp với các tham số đầu vào:

- : Dirichlet prior on (theta)

-: Dirichlet prior on

- : là phân bố chủ đề trên tài liệu thứ ***m*** (tham số mức tài liệu). biểu diễn tham số cho đây là xác suất để topic ***z*** thuộc về tài liệu thứ ***m*** (thành phần trộn topic cho tài liệu ***m***). Một tỷ lệ cho mỗi tài liệu.

* Nếu giá trị của cao điều đó có nghĩa rằng tài liệu thứ ***m*** được tạo nên từ hầu hết tất cả các topic, không phải chỉ riêng một topic nào.
* Nếu giá trị của thấp, tài liệu thứ ***m*** có thể chỉ được tạo nên từ một số ít, rất ít hoặc chỉ một topic mà thôi.

- : là phân phối của các từ được sinh từ topic . biểu diễn tham số cho đây là xác suất xảy ra biến cố ***t*** là một từ bất kì có khả năng thuộc về topic ***z*** thứ ***k***, thành phần trộn của topic ***k***.

* Khi giá trị của cao có nghĩa rằng mỗi chủ đề có khả năng chứa một tập hợp hầu hết các từ trong tập từ vựng hiện tại, và các từ này có xác suất gần như tương đương nhau.
* Khi giá của thấp, thì nhiều khả năng topic ***z*** đang được nhắc đến, chỉ là tập hợp của một ít, hoặc một vài từ trong tập từ vựng.

🡪 Như vậy, nếu các giá trị và quá cao, rất có thể có một sự lặp lại của các chủ đề, cũng như các chủ đề lại chứa nhiều từ giống nhau.

- : topic index (word ***n*** của văn bản ***m***).

-: word ***n*** của văn bản ***m*** được chỉ ra bởi (word-level variable, biến cấp độ từ).

- **M**: số lượng các tài liệu đầu vào.

- **Nm**: số lượng các từ trong tài liệu thứ m (độ dài văn bản).

- **K**: số lượng topic ẩn.

- *Dir* và *Mult*: là các phân bố Dirichlet và phân bố Multinominal đa thức.

### Ước lượng tham số và suy luận

Ước lượng tham số cho mô hình LDA bằng tối ưu hóa một cách trực tiếp và chính xác xác suất của toàn bộ tập dữ liệu là khó có thể thực hiện. Một giải pháp được đề ra là sử dụng phương pháp ước lượng xấp xỉ như phương pháp biến phân [8] và lấy mẫu Gibbs [17].

LDA sẽ được tính toán bằng việc khởi tạo các biến số cần thiết, khởi tạo chúng một giá trị ngẫu nhiên, sau đó thực hiện các vòng lặp và trong mỗi vòng lặp các “chủ đề” sẽ được lấy mẫu cho mỗi từ trong tập dữ liệu. Sau khi thực hiện các vòng lặp, các biến đếm có thể được sử dụng để tính toán “phân phối ẩn” và**.**

Các giá trị biến số cần được yêu cầu bao gồm: số lượng từ tạo nên chủ đề *k* trong tập tài liệu ***d***; vàsố lần mà từ được gán cho “chủ đề”Tuy nhiên, để đơn giản và hiệu quả thì các giá trị tổng số lần mà từ bất kỳ được gán cho “chủ đề” Cuối cùng, ngoài các biến rõ ràng đại diện cho corpus – tập toàn bộ dữ liệu, thì cũng cần một mảng**.** Mảng này sẽ chứa danh sách các “chủ đề” hiện tại đang được gán cho mỗi từ ***N*** trong corpus. Chúng ta sẽ thực hiện lần lược các bước để có thể suy luận các chủ đề ***z*** và các từ ***w*** thuộc về từng chủ đề một:

* Bước 1: Duyệt qua từng tài liệu một, và thực hiện việc gán ngẫu nhiên mỗi từ trong từng tài liệu cho một trong các chủ đề của chủ đề. Chú ý rằng việc gán ngẫu nhiên như vậy cung cấp các chủ đề đại diện cho tập dữ liệu và phân phối của các từ vào các chủ đề tương ứng đó.
* Bước 2: Tiến hành cải thiện các kết quả bằng cách, duyệt qua từng tài liệu
  + Bước 2a: Duyệt qua từng từtrong
  + Bước 2b: Với mỗi topic tính toán hai giá trị:
    - * là xác suất của các từ trong tài liệuhiện tại đang được gán cho topic
      * là khả năng mà từthuộc về topictrên toàn bộ tập dữ liệu.
  + Bước 2c: Tiến hành gán từvào một topic mới dựa vào xác suất:(theo mô hình sinh của chúng ta, điều này về cơ bản chính là xác suất mà topic được tạo ra từ từ , do đó nó cũng có nghĩa là chúng ta đã resample (lấy mẫu lại) chủ đề mà từ hiện tại đang thuộc về với xác suất này). Nói một cách khác, ở bước này chúng ta đang giả định rằng tất cả các topic đều đã được gán bởi một tập từ vựng nào đó, ngoại trừ từ hiện tại đang xét là đã đúng. Sau đó, chúng ta sẽ cập nhật lại việc gán của từ hiện tại vào topic, bằng cách lặp lại quá trình trên.
* Bước 3: Thực hiện việc lặp lại các quá trình trên nhiều lần, các topic sẽ dần dần đạt được một trạng thái ổn định, và việc tạo ra topic – document sẽ hoàn thành. Vì vậy, sử dụng việc những tập dữ liệu trên để có thể ước tính tập hợp topic nào cho mỗi document (bằng cách đếm tỉ lệ những từ đã được gán cho một topic trong document ấy) và những từ liên quan đến mỗi topic (bằng cách đếm tổng thể tỉ lệ của những từ đã được gán cho mỗi topic).

Dưới đây là mã giải chi tiết của việc lấy mẫu Gibb theo các tác giả trong tài liệu [17].

**Input***:* words

**Ouput***:* topicassignmentsandcountsvà

**begin**

randomlyinitializeandincrementcounters

**foreach** *iteration* **do**

**fordo**

**fordo**

**end**

**end**

**end**

**return** *z,*

**end**

## Phân tích và đánh giá cảm xúc

### Định nghĩa quan điểm

Theo Bo Pang và Lillian, 2008 [7], thuật ngữ **“Khai phá quan điểm”** (Opinion Minning) xuất hiện lần đầu tiên trong bài báo cáo của Kushal Dave và đồng sự [12] được công bố ở hội nghị WWW năm 2003. Thuật ngữ này có thể giải thích theo nhiều cách khác nhau, liên quan chặt chẽ với tìm kiếm và trích chọn thông tin trên Web. Theo Kushal Dave và đồng sự, công cụ khai phá quan điểm lý tưởng sẽ “xử lý một tập các kết quả tìm kiếm cho một đối tượng nhất định, tạo ra một danh sách các thuộc tính (chất lượng, đặc trưng…) và tổng hợp quan điểm cho mỗi thuộc tính đó (xấu, tốt,…)”. Tuy nhiên, đến khi nghiên cứu của Bing Liu [6] được công bố năm 2006 thì thuật ngữ khai phá quan điểm mới được đưa ra một cách rõ ràng hơn.

Cũng theo Bing Liu, các quan điểm có thể là thể hiện về bất cứ điều gì, ví dụ như một sản phẩm, một cá nhân, một tổ chức, một chủ đề…Ông coi các thực thể được nhận xét là đối tượng (Object). Đối tượng này gồm một tập hợp các thành phần (components) và một tập các thuộc tính (attributes). Theo đó, tác giả đã đưa ra được một số khái niệm trong khai phá quan điểm:

Một quan điểm bao gồm 2 thành phần chính là: một mục tiêu *g* và một cảm nhận *s* nào đó cho mục tiêu *g.*

Biểu diễn quan điểm: (*g, s*)

Trong đó *g* có thể là một thực thể hay một khía cạnh của thực thể bất kì nào đó được đánh giá bằng một quan điểm, *s* là một cảm xúc tích cực, tiêu cực hoặc trung tập, hoặc cũng có thể là một điểm số đánh giá nào đó thể hiện mức độ cảm xúc (Ví dụ: 1 đến 5 sao). Sự tích cực, tiêu cực và trung lập được gọi là những định hướng cảm xúc/quan điểm. [6]

Ví dụ: *Tình yêu thời sinh viên thật sự rất đẹp*.

Ở ví dụ đơn giản này mục tiêu *g*  là “*Tình yêu thời sinh viên*” và cảm xúc *s* là “*rất đẹp*”.

Xét ví dụ: *“Bữa tớ cũng cãi nhau với thằng học luật vì nó quá ngu lúc thú tội, còn thằng bạn tớ thì nói nó làm vậy là thông minh và tốt bụng mà, ức chế”. – bình luận vào ngày 10/11/2015.*

Ở ví dụ này, có 2 người đưa ra 2 quan điểm về một mục tiêu. Quan điểm từ 2 người được gọi là *người giữ quan điểm (opinion holders)* (Kim và Hovy, 2004; Wiebe, Wilson và Cardie, 2005).

Thời gian của bình luận trong ví dụ trên là “*10/11/2015*”, thời gian của quan điểm khá quan trọng trong thực tế, vì chúng ta cần biết quan điểm thay đổi như thế nào và xu hướng của quan điểm là gì.

Từ những ý trên, chúng ta sẽ định nghĩa quan điểm là một bộ tứ.

Biểu diễn quan điểm: *(g, s, h, t),*

Trong đó *g* là mục tiêu của quan điểm (hoặc cảm xúc), *s* là cảm xúc về mục tiêu, *h* là người đưa ra quan điểm/cảm xúc (opinion holder) và *t* là thời gian khi quan điểm được thể hiện.

### Bài toán phân tích cảm xúc

Quan điểm là trung tâm trong hầu hết các hoạt động của con người, bởi vì nó có ảnh hưởng nhiều nhất đến cách hành xử của con người. Bất cứ khi nào chúng ta cần ra quyết định, chúng ta đều muốn biết quan điểm của những người khác. Trong thực tế, các doanh nghiệp và tổ chức luôn muốn biết những quan điểm của khác hàng hoặc quan điểm của cộng đồng về sản phẩm và dịch vụ của họ. Bản thân mỗi khách hàng cũng muốn biết quan điểm của những người khác đã sử dụng sản phẩm nào đó trước khi mua sản phẩm. Tương tự như vậy, khi một người muốn bỏ phiếu bầu cử cho một ai đó thì họ cũng muốn biết quan điểm của những người khác về các ứng cử viên. Khi một cá nhân cần tham khảo ý kiến về một điều gì đó, họ thường hỏi gia đình, bạn bè. Khi một doanh nghiệp hoặc tổ chức cần tham khảo ý kiến của khách hàng hoặc cộng đồng thì họ làm các khảo sát, thu thập ý kiến… Biết được các ý kiến/quan điểm của khách hàng, cộng đồng là một điều rất quan trọng và đem lại lợi thế to lớn trong việc quảng bá sản phẩm, quan hệ công chúng hay trong việc quản lý một tập thể lớn. Ví dụ: Các phòng ban quản lý cần biết thông tin về tình hình và những vấn đề quan tâm của tập thể sinh viên trong một trường đại học hay trong kí túc xá để có những hành động, điều chỉnh kịp thời.

Với sự bùng nổ của các phương tiện thông tin xã hội như: mạng xã hội, diễn dàn, blog, các trang đánh giá sản phẩm… thì thông tin về quan điểm của khách hàng, của cộng đồng tăng lên một cách nhanh chóng. Các cá nhân, doanh nghiệp hay tổ chức thậm chí không cần thiết phải hỏi ý kiến của người khác hoặc làm khảo sát, thăm dò ý kiến mà vẫn có được các thông tin về quan điểm và ý kiến của cộng đồng/khách hàng. Tuy nhiên, để có được các thông tin này một cách khách quan là điều không dễ dàng, vì số lượng các đánh giá, bình luận quá lớn và gia tăng liên tục làm cho việc đọc các hết các thông tin, ý kiến, bình luận là không thể thực hiện được. Đó là lý do việc đánh giá cảm xúc và phân lớp quan điểm một cách tự động trở nên rất cần thiết.

Những thông tin trên internet rất đa dạng và bao gồm rất nhiều lĩnh vực, cho nên những ứng dụng phân tích cảm xúc cũng phân bố trên hầu hết các lĩnh vực từ những người sử dụng sản phẩm, dịch vụ, chăm sóc sức khỏe, dịch vụ tài chính, sự kiện xã hội, bầu cử…

#### Các bài toán trong phân tích cảm xúc

Phân tích cảm xúc hay còn gọi là phân lớp nhận định có ba bài toán điển hình nhất đó là:

* Phân lớp quan điểm: Với bài toán này có thể coi khai phá quan điểm như bài toán phân lớp văn bản. Ở đây bài toán sẽ là phân lớp một văn bản đánh giá là tích cực hay tiêu cực. Ví dụ: với một đánh giá về một sản phẩm, hệ thống xác định xem nhận xét về sản phẩm là tốt hay xấu. Phân lớp này thường là phân lớp ở mức tài liệu. Thông tin được phát hiện không mô tả chi tiết về những gì mọi người thích hay không thích.
* Khai phá và tổng hợp quan điểm dựa trên đặc trưng: Bài toán đi sâu vào phát hiện quan điểm ở mức câu, tức là tìm hiểu các khía cạnh của đối tượng mà người bày tỏ quan điểm là thích hay không thích. Một đối tượng có thể là sản phẩm, một dịch vụ, một sự việc, một chủ đề hay một cá nhân, tổ chức nào đó…Ví dụ, trong một bài đánh giá về sản phẩm, bài toán phải xác định được các đặc trưng mà người dùng bình luận, đánh giá đó là gì, và xác định các ý kiến là tích cực hay tiêu cực. Ví dụ: Trong một đánh giá: “Pin của laptop này mau hết pin quá”. Thì bài toán phải xác định rõ, chủ thể được đánh giá là “Pin của laptop”, và quan điểm ở đây là “tiêu cực”.
* Khai phá quan hệ (so sánh câu): So sánh là một loại đánh giá khác, tức là so sánh trực tiếp một đối tượng đối với một hoặc nhiều đối tượng khác tương tự. Ví dụ: như so sánh, đánh giá về hai lớp học:

“Lớp của cô A có vẻ thú vị hơn lớp của thầy B nhiều”.

Ở đây, khóa luận xin đề cập tới bài toán thứ nhất là bài toán phân lớp quan điểm: coi khai phá quan điểm như là phân lớp văn bản, coi mỗi quan điểm là một văn bản và quá trình phân lớp quan điểm chính là phân lớp văn bản. Các quan điểm sẽ được phân vào 3 lớp tích cực (tốt), tiêu cực (xấu) và lớp trung tính (không bộc lộ cảm xúc). Mục tiêu chủ đạo là nhanh chóng xác định quan điểm đánh giá về một đối tượng là tốt hay xấu.

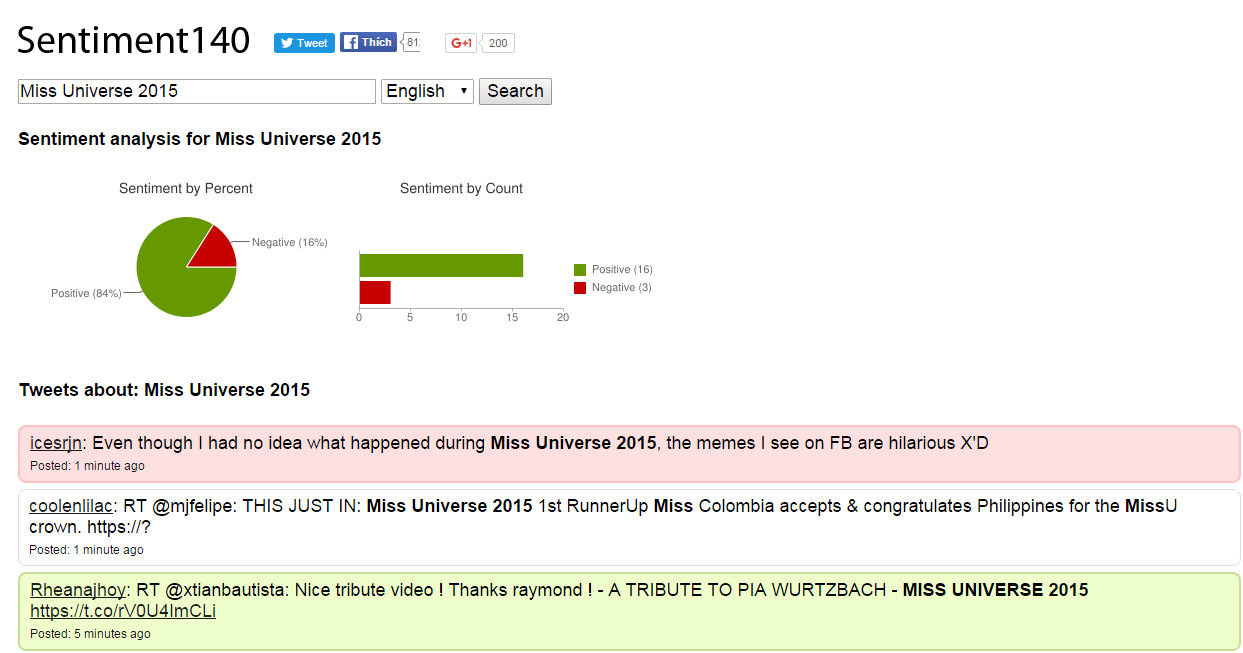
### Các nghiên cứu về và công cụ cho việc phân tích cảm xúc đã có

Hiện nay cũng đã có nhiều bài báo nghiên cứu hướng ứng dụng về đánh giá cảm xúc. Ví dụ, (Liu et al., 2007), nói về một mô hình cảm xúc đã được đề xuất để tăng hiệu quả bán hàng. (McGlohon, Glance and Reiter, 2010), đã sử dụng những phản hồi để đánh giá phân lớp cho sản phẩm. (Hong and Skiena, 2010) nghiên cứu về những mối quan hệ giữa “NFL betting line” và những quan điểm công khai trên blog và Twitter. (O’Connor et al., 2010), những ý kiến trên Twitter được liên kết với quan điểm của những cuộc bầu cử công khai. (Tumasjan et al., 2010) những ý kiến trên Twitter cũng được áp dụng cho việc dự đoán kết quả bầu cử, …

Về các công cụ phân tích cảm xúc, hiện tại đã có nhiều nghiên cứu cung cấp các phương pháp và công cụ được sử dụng để phân tích và đánh giá cảm xúc. Các công cụ được sử dụng để phát hiện tính phân cực của cảm xúc (bao gồm cảm xúc tích cực và cảm xúc tiêu cực) của một tin nhắn hoặc một đoạn văn dựa trên các từ trong đó, đặc biệt là dựa vào các biểu tượng cảm xúc. Các biểu tượng cảm xúc này mô phỏng các hình thái của khuôn mặt từ đó cho biết được cảm xúc hiện tại là buồn bã hay vui vẻ. Mặc dù thực tế chúng ta có hàng loạt các sắc thái cảm xúc khác nhau trên khuôn mặt, các “emoticons” này cũng phần nào hình tượng hóa cảm xúc của người viết. Để diễn đạt và hiểu ý nghĩa các biểu tượng cảm xúc chúng ta cần có một từ điển phổ biến giải thích chúng được lấy từ <http://cool-smileys.com/text-emoticons> và trong khuôn khổ của luận văn, chúng tôi sử dụng   
[http://www.symbols-n-emoticons.com/p/facebook  
-emoticons-list.html](http://www.symbols-n-emoticons.com/p/facebook-emoticons-list.html) cho việc phân tích cảm xúc dựa trên các “emoticons” phổ biến của Facebook.

Do đó, các biểu tượng cảm xúc thường được các công cụ sử dụng kết hợp với các kỹ thuật khác để xây dựng một tập dữ liệu training cho các công cụ supervised machine learning. Một phương pháp khác nữa là Linguistic Inquiry and Word Count [14] cho phép việc phân tích trạng thái cảm xúc không chỉ dựa trên cảm xúc tích cực hay tiêu cực mà còn dựa vào tình cảm, nhận thức, và cấu trúc của một văn bản bằng cách sử dụng một từ điển chứa các từ và danh mục phân loại của các từ đó. Ví dụ, từ “agree” sẽ thuộc về các loại từ sau: assent (sự tán thành), affective (tình cảm), positive emotion (cảm xúc tích cực) và cognitive process (thuộc về quá trình nhận thức). Một phần mềm điển hình thuộc dạng như vậy là LIWC (<http://liwc.wpengine.com/>).

Công cụ tiếp theo là SentiStrength (<http://sentistrength.wlv.ac.uk/>) đây là một trong những công cụ phân tích cảm xúc phổ biển nhất [18]. Nó sử dụng một bộ từ vựng cảm xúc để tính toán trọng số tích cực và tiêu cực của một câu (giá trị từ: -1, không tiêu cực đến -5 rất tiêu cực và 1 không tích cự**c** đến 5 rất tích cực). Một trong những nghiên cứu đánh giá cảm xúc đã được ứng dụng và có độ chính xác khá cao nữa đó là trang web Twitter Sentiment (<http://www.sentiment140.com>). Với đầu vào là tên thực thể người dùng cần nắm quan điểm, hệ thống đưa ra tổng hợp các bình luận của người dùng Twitter về thực thể đó, đồng thời đưa ra tỷ lệ tích cực/tiêu cực các quan điểm về thực thể đó. Kết quả tổng hợp được đưa ra kèm theo thời điểm và phần thống kê tỉ lệ phần trăm Pos/Neg của các quan điểm. Thêm vào đó, hệ thống còn cho phép người dùng đánh giá lại kết quả phân tích của hệ thống, điều này giúp hệ thống cải thiện được kết quả cao hơn cho lần sau.



Hình 2.3: Trang web Twitter Sentiment với từ khóa “Miss Universe 2015”

Tuy nhiên các bộ công cụ này hầu hết hỗ trợ cho Tiếng Anh hoặc các thứ tiếng khác mà chưa thật sự có công cụ nào hỗ trợ tối đa cho Tiếng Việt. Trong phần sau của luận văn, sẽ giới thiệu hai công cụ có hỗ trợ Tiếng Việt mà nhóm sử dụng cho khả năng phân tích cảm xúc của luận văn.

### Các phương pháp tiếp cận cho việc phân tích cảm xúc

Dựa vào các công trình nghiên cứu đã có cũng như các phương pháp phân tích cảm xúc, có thể tóm lược lại các phương pháp tiếp cận như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **PHƯƠNG PHÁP TIẾP CẬN** | | **TÍNH NĂNG &**  **KĨ THUẬT** | **ƯU ĐIỂM VÀ HẠN CHẾ** |
| **Machine Learning** | Bayesian Network  Naïve Bayes Classification  Maximum Entropy  Neutral Networks  Support Vector Machine | Tần số xuất hiện của các từ  Từ loại (danh từ, tính từ)  Các từ phủ định  Từ và cụm từ chỉ quan điểm | **ƯU ĐIỂM**: có khả năng thích ứng để tạo ra dữ liệu học phù hợp với mục đích và bối cảnh cụ thể.  **HẠN CHẾ**: khả năng ứng dụng để xử lý với dữ liệu mới còn thấp, vì nó cần phải có dữ liệu sẵn có, và quá trình này rất tốn chi phí và thời gian |
| **Lexicon Base** | Dựa trên tập từ vựng cảm xúc | Sự đánh giá thủ công  Dựa trên tập dữ liệu lớn  Dựa trên các tập từ điển sẵn có | **ƯU ĐIỂM**: khả năng bao phủ các trường hợp rộng hơn.  **HẠN CHẾ**: chỉ có một tập hữu hạn các từ được xác định sẵn với các trọng số cũng đã được tính toán sẵn. |
| **Hybrid** | Kết hợp phương pháp Lexicon và Machine Learning | Sử dụng phương pháp Lexicon để phân lớp dữ liệu học. Các từ chứa cảm xúc sẽ là các đặc trưng, được sử dụng cho cách tiếp cận Machine Learning | **ƯU ĐIỂM**: sử dụng kết hợp khả năng phân lớp của Lexicon và sự đánh giá của Machine Learning  **HẠN CHẾ**: cần phải có quá trình xem xét và đánh giá lại kết quả |

Bảng 2.1: So sánh các giải pháp đã có cho việc phân tích cảm xúc

Cách tiếp cận Machine learning (tiếp cận bằng phương pháp máy học), sẽ sử dụng các kỹ thuật phân loại để tiến hành phân loại văn bản; quá trình này sẽ cần đến hai bộ dữ liệu độc lập nhau: dữ liệu training (huấn luyện), và dữ liệu test (kiểm tra). Tập dữ liệu huấn luyện, được sử dụng nhằm giúp cho mô hình máy học nhận biết các đặc điểm khác biệt giữa các tài liệu với nhau, trong khi đó tập dữ liệu kiểm tra được sử dụng để đánh giá lại kết quả phân loại “chính xác” như thế nào. Các tính năng mà phương pháp Machine learning (phương pháp máy học) áp dụng cho phân tích cảm xúc là:

* Sự xuất hiện của từ và tần số xuất hiện của từ đó: bao gồm cả uni-grams hoặc n-grams.
* Thông tin về part of speech (loại từ vựng): sử dụng thông tin này để định hướng ý nghĩa của từ, và sử dụng nó để lựa chọn đặc trưng.
* Từ ngữ phủ định: có khả năng đảo ngược quan điểm cảm xúc của từ/cụm từ.

Các phương pháp tiếp cận dựa trên các từ vựng cảm xúc, sử dụng các từ điển với các từ cảm xúc kết hợp với trọng số của chúng để đưa ra xác định về tính phân cực. Có ba kĩ thuật để xây dựng từ điển cảm xúc lexicon: xây dựng thủ công, xây dựng dựa trên phương pháp corpus (tập dữ liệu lớn), và phương pháp dựa trên từ điển. Phương pháp xây dựng thủ công là một nhiệm vụ khó khăn và rất tốn thời gian. Phương pháp dựa trên corpus có thể tạo ra tập từ quan điểm với độ chính xác tương đối cao. Cuối cùng, trong các kỹ thuật dựa vào từ điển, ý tưởng là việc thu thập một tập nhỏ dữ liệu bằng tay những từ phổ biến và thường được sử dụng. Sau đó phát triển bộ từ điển này, bằng cách tìm kiếm trong từ điển WordNet các từ đồng nghĩa và trái nghĩa của nó.

Cuối cùng là trong phương pháp Hybrid – phương pháp kết hợp cả hai giải pháp sử dụng Machine learning (máy học) và phương pháp tiếp cận dựa trên bộ từ điển cảm xúc để cải thiện hiệu suất, và độ chính xác cho khả năng phân loại. Có một số ưu điểm và hạn chế trong việc sử dụng các phương pháp tiếp cận khác nhau, dựa trên mục đích của việc phân tích là gì. Cũng theo các tác giả của [2] cung cấp một cái nhìn tổng quan nhất.

Ưu điểm chính của phương pháp tiếp cận máy học là khả năng thích ứng và tạo ra mô hình đào tạo cho các mục đích và bối cảnh cụ thể, trong khi đó khó khăn, đó là các dữ liệu thực tế có thể nằm ngoài dữ liệu được huấn luyện. Hơn nữa, các mô hình máy học này có khả năng thích ứng kém giữa các lĩnh vực hoặc các thể loại văn bản khác nhau, vì nó phải dựa trên những đặc trưng mà tập dữ liệu huấn luyện cung cấp.

Phương pháp tiếp cận dựa trên tập từ vựng lexicon có lợi thế là khả năng bao quát các từ chỉ trạng thái cảm xúc là cao hơn, có độ bao phủ dữ liệu lớn hơn. Tuy nhiên các phương pháp này có hai hạn chế chính. Thứ nhất, số lượng từ trong lexicon là hữu hạn, khó có thể giải quyết trong môi trường có nhiều từ vựng mới chưa được giải nghĩa. Thứ hai, các từ vựng của lexicon có xu hướng cố định trạng thái cảm xúc của các từ vào các mức độ nhất định, bất kể trong các trường hợp hay ngữ cảnh khác chúng cũng không thay đổi.

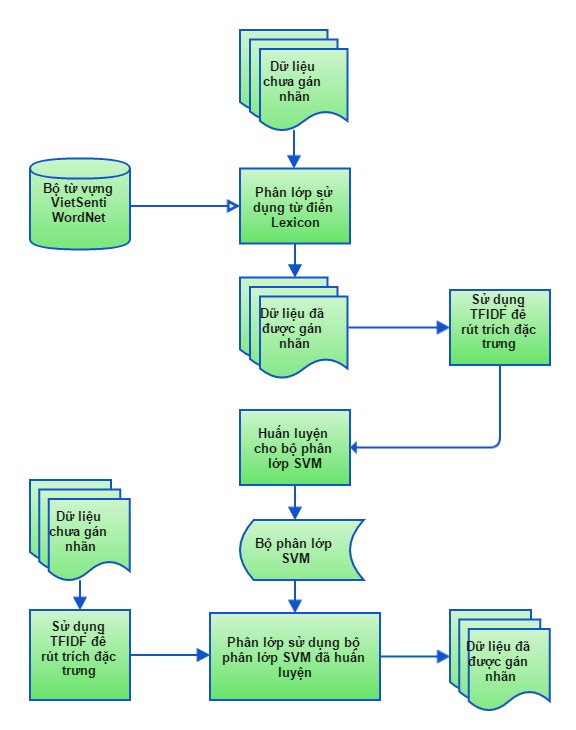
Ưu điểm chính của phương pháp lai – phương pháp kết hợp lexicon/machine learning, là việc phát hiện và đo lường tính phân cực cảm xúc ở mức khái niệm và mức độ nhạy cảm thấp hơn trong phạm vi một chủ đề nhất định. Trong khi các hạn chế lớn nhất là các đánh giá thường sẽ bị nhiễu bởi những từ trung hòa, không mang ý nghĩa cảm xúc.

### Đánh giá cảm xúc dựa vào từ vựng (Lexicon) và vấn đề gặp phải

Những cảm xúc được xác đinh chủ yếu từ những từ vựng cảm xúc (sentiment words) cũng được gọi là từ vựng quan điểm (opinion words). Có nhiều từ phồ biến thường được dùng để biểu hiện cảm xúc như: *tốt*, *tuyệt*, *tuyệt vời* là những từ cảm xúc tích cực và *tồi, tội nghiệp, khinh khủng* là những từ cảm xúc tiêu cực. Ngoài những từ cảm xúc còn có những cụm từ hay thành ngữ nói về cảm xúc, ví dụ: *gừng càng già càng cay, không gì là không thể…* Mặc dù những từ vựng, cụm từ, thành ngữ là cần thiết và rất quan trọng trong việc đánh giá cảm xúc, nhưng chỉ sử dụng chúng thì hoàn toàn chưa đủ để có thể đánh giá cảm xúc. Vì một số lý do như sau:

* Một từ tích cực hoặc tiêu cực có thể có ý nghĩa ngược lại trong những lĩnh vực khác nhau. Ví dụ: *Những lưỡi dao này rất* ***cứng*** *và sắc bén/ Bàn phím của chiếc điện thoại này quá* ***cứng*** *nên thao thác khó khăn.*
* Một câu có chứa các từ cảm xúc có thể không thể hiện cảm xúc nào cả. Ví dụ một câu hỏi: *Bạn có thể chỉ có mình cái máy quay nào tốt hơn không?*  hoặc *Nếu hôm nay tôi tìm được một chiếc xe tốt, tôi sẽ mua nó ngay.* Cả hai câu này đều chứa từ “*tốt*” nhưng cả hai câu đều không thể hiện cảm xúc tích cực nào cả.
* Những câu mang ý nghĩa chế nhạo cho dù có chứa từ cảm xúc hay không đều rất khó để đánh giá đúng cảm xúc. Ví dụ: *Chiếc xe này tuyệt thật, cứ chạy được tầm 10km là tắt máy một lần.*
* Nhiều câu không chứa từ cảm xúc nhưng vẫn thể hiện cảm xúc tích cực hoặc tiêu cực. Ví dụ: *Cái tủ lạnh này phát ra tiếng động quá lớn khi hoạt động.*

### Đề xuất về mô hình phân tích cảm xúc kết hợp phương pháp phân lớp Lexicon và phương pháp SVM.



Hình 2.4: Mô hình phân lớp cảm xúc đề xuất

***Các bước chính trong mô hình:***

**Bước 1:** **Phân lớp bằng Lexicon sử dụng bộ từ điển cảm xúc VietSentiWordNet**

* Ở bước này hệ thống tiến hành so khớp mỗi tài liệu chưa được gán nhãn với tập từ vựng cảm xúc từ đó tính được điểm số của tài liệu. Nếu điểm số > 0, tài liệu sẽ được gán nhãn là 1 (tích cực), nếu điểm số < 0, tài liệu được gán nhãn là -1 (tiêu cực) và trường hợp còn lại sẽ được gán nhãn là 0 (trung lập)
* Tài liệu sau khi được gán nhãn sẽ được lưu ra bộ nhớ ngoài theo định dạng mỗi dòng là: Mã tài liệu <tab> Nhãn phân lớp <tab> Nội dung tài liệu.

**Bước 2: Rút trích đặc trưng của tài liệu**

* Ở bước này hệ thống sử dụng một thực thi của thuật toán TFIDF để rút trích đặc trưng của tài liệu tạo thành vector đặc trưng.
* Những đặc trưng của tài liệu sau đó được chuyển đổi thành định dạng dữ liệu mà SVM hỗ trợ để tiến hành huấn luyện cho SVM.

**Bước 3: Huấn luyện cho bộ phân lớp SVM**

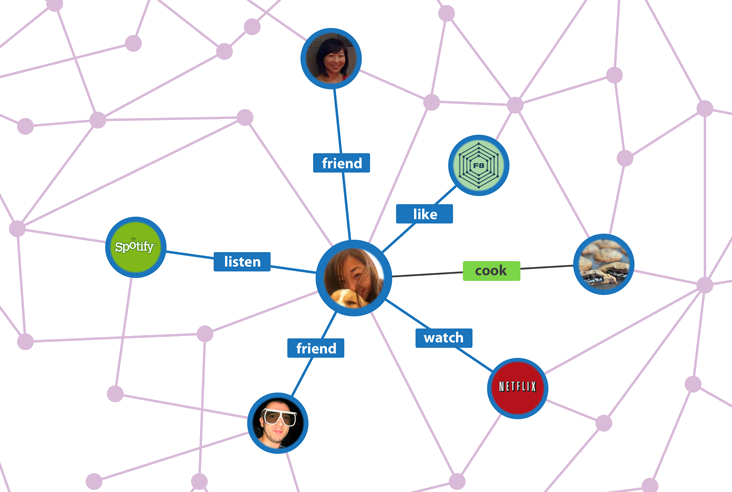
* Hệ thống sử dụng dữ liệu đầu vào từ bước rút trích đặc trưng, dữ liệu được chia ngẫu nhiên thành 2 tập dữ liệu khác nhau theo tỉ lệ 60% và 40% tương ứng là dữ liệu huấn luyện và dữ liệu kiểm tra. Tập huấn luyện được sử dụng để huấn luyện cho bộ phân lớp SVM. Sau khi huấn luyện hệ thống sử dụng tập kiểm tra để kiểm tra kết quả huấn luyện.

**Bước 4: Phân lớp cho dữ liệu sử dụng SVM**

* Dữ liệu đầu vào sẽ được phân lớp bằng bộ phân lớp SVM đã được huấn luyện.

## Giới thiệu thư viện hỗ trợ lấy dữ liệu từ Facebook - RestFB

Ngày nay, đa số các trang web đều được tích hợp với các trang mạng xã hội như Facebook hay Twitter. Các trang mạng xã hội này chỉ phát huy được tối đa khả năng của nó khi mà nó cung cấp các API, cho phép các trang web khác tương tác với nó. Với Facebook, thì hiện tại họ đang cung cấp được các định nghĩa Graph API mới thuận tiện và rất đầy đủ, được công khai đầy đủ ở <http://developers.facebook.com/docs/api/>. Facebook coi các mối quan hệ giữa các thực thể như là một “Đồ thị xã hội” (Social Graph).



Hình 2.5 : Mô tả Social Graph

Facebook Graph APIlà cách chủ yếu để lấy dữ liệu vào và ra khỏi đồ thị xã hội này. Đó là một HTTP API dựa trên một mức độ thấp, mà nhờ đó bạn có thể sử dụng để truy vấn dữ liệu, đăng tải những bài đăng, hình ảnh mới và các tác vụ khác mà các ứng dụng có thể cần yêu cầu để thực hiện. Facebook Graph APIhoạt động dựa trên giao thức HTTP, do đó, nó có khả năng làm việc với bất kỳ ngôn ngữ nào có thư viện HTTP như cURL, urllib... Về cơ bản các đại diện thông tin trên Facebook bao gồm:

**Node (**nút**):** Một cách cơ bản, đây là các thực thể, những thành viên, một hình ảnh, một trang, một nhận xét trên Facebook.

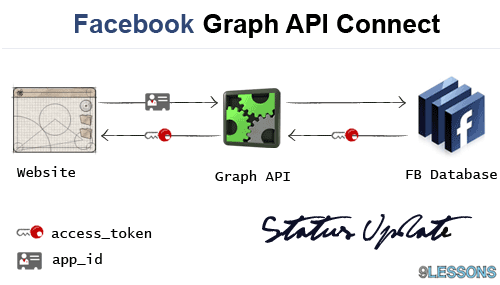
**Edge (**cạnh**):** Là các kết nối giữa các thực thể với nhau, các kết nối này có thể là giữa các hình ảnh và trang chứa hình ảnh đó, một ghi chú, và một hình ảnh trong ghi chú đó.

**Field** (trường/lĩnh vực): Thông tin về các thực thể, chẳng hạn như ngày sinh nhật, tên của một fanpage.

Để có thể thao tác và làm việc với tập dữ liệu cực kì lớn của Facebook, khóa luận sử dụng một trong những giải pháp được xem là tối ưu và linh họat, đó là công cụ RestFB API. RestFB là một API gọn nhẹ của Facebook Gragh API (nền tảng công nghệ dữ liệu của Facebook). RestFB được viết bởi Mark Allen và sau đó được bảo trì và phát triển bởi Norbert Bartels. Phiên bản mới nhất của RestFB là 1.18.0 được giới thiệu vào ngày 21/12/2015. Đây là một phần mềm mã nguồn mở giúp kết nối Facebook với các kênh thông tin web khác cũng như với các ứng dụng của bên thứ ba. RestFB được công khai dựa vào giấy phép mã nguồn mở MIT. Các thông tin về API, cấu hình, mã nguồn cũng như các ví dụ cụ thể được chia sẻ đầy đủ tại <http://restfb.com/>.

Các tính năng của RestFB API bao gồm những API được Facebook Graph API công khai, cùng với đó là khả năng mở rộng tối đa, mạnh mẽ với những thay đổi thường xuyên của Facebook API, được cấu hình để có thể khai thác tối đa các siêu dữ liệu theo một cách cực kỳ đơn giản. Một đặc điểm quan trọng của RestFb là bạn có thể dễ dàng lấy được những thông tin cơ bản của người dùng Facebook, thông qua các phương thức được cài đặt với ngôn ngữ Java. RestFB giúp trích xuất danh sách bạn bè Facebook, các bài đăng và bình luận. Tạo kết nối với Facebook thông qua mã truy cập (Access token), lấy thông tin theo kết nối về thông tin người dùng, trang cá nhận, các trang công khai…, hỗ trợ tìm kiếm thông tin từ Facebook, truy vấn dữ liệu Facebook bằng FQL, hỗ trợ đăng/xóa bài lên Facebook của người dùng, quản lý các trang mà người dùng sở hữu. Hỗ trợ quản lý lỗi và tăng hiệu suất của các hàm.

Để có thể thực hiện được các truy vấn đến cơ sở dữ liệu của Facebook thông qua Facebook Graph API**,** Facebook đều yêu cầu một mã “Access Tokens” đây là mã cho phép gởi yêu cầu trích xuất thông tin đến server của Facebook. Nếu bạn đang đăng nhập vào một tài khoản Facebook, thì giá trị mặc định hiển thị là cho tài khoản đó.

****

Hình 2.6: Mô tả kết nối với cơ sở dữ liệu của Facebook

Một số ứng dụng đang sử dụng RestFB như: Game Angry Bird, IBM, Hubspot, DwinQ, Friend2Friend... [11]

## Giới thiệu công cụ tách từ Tiếng Việt – vnTokenizer

Tách từ là một quá trình xử lý nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể hiểu đơn giản rằng tách từ là quá trình xác định các từ đơn, từ ghép… có trong câu. Đối với xử lý ngôn ngữ, để có thể xác định cấu trúc ngữ pháp của câu, xác định từ loại của một từ trong câu, yêu cầu nhất thiết đặt ra là phải xác định được đâu là từ trong câu. Vấn đề này tưởng chừng đơn giản với con người nhưng đối với máy tính, đây là bài toán rất khó giải quyết.

Chính vì lý do đó tách từ được xem là bước xử lý quan trọng đối với các hệ thống Xử Lý Ngôn Ngữ Tự Nhiên, đặc biệt là đối với các ngôn ngữ thuộc vùng Đông Á theo loại hình ngôn ngữ đơn lập, ví dụ: tiếng Trung Quốc, tiếng Nhật, tiếng Thái, và tiếng Việt. Với các ngôn ngữ thuộc loại hình này, ranh giới từ không chỉ đơn giản là những khoảng trắng như trong các ngôn ngữ thuộc loại hình hòa kết như tiếng Anh…, mà có sự liên hệ chặt chẽ giữa các tiếng với nhau, một từ có thể cấu tạo bởi một hoặc nhiều tiếng. Vì vậy đối với các ngôn ngữ thuộc vùng Đông Á, vấn đề của bài toán tách từ là khử được sự nhập nhằng trong ranh giới từ. [19]

Trong khuôn khổ khóa luận, chúng tôi sử dụng công cụ tách từ dành cho Tiếng Việt – vnTokenizer. Công cụ này tách từ đúng ở độ chính xác và hồi quy trong khoảng 96% đến 98%.

## Giới thiệu về kho ngữ liệu WordNet

Bài toán khai phá quan điểm sử dụng nhiều kết quả của xử lý ngôn ngữ tự nhiên, trong đó, một kho ngữ liệu chuẩn là cần thiết. **WordNet** là một kho ngữ liệu tiếng Anh, các từ tiếng Anh được nhóm thành các bộ từ đồng nghĩa được gọi là các *synset,* cung cấp các định nghĩa chung và ngắn gọn, đồng thời, ghi lại giá trị quan hệ ngữ nghĩa học giữa các bộ đồng nghĩa. **WordNet** được Cognitive Science Laboratory of Princeton University dưới sự chỉ đạo của Giáo sư George A.Miller tạo ra và duy trì. Được phát triển từ năm 1985, đến nay phiên bản mới nhất là 3.0 chứa 155,287 từ được tổ chức thành 117,659 synsets [20], kích cỡ khoảng 12MB. Dữ liệu và bộ công cụ **WordNet** có thể tải và sử dụng miễn phí tại trang **WordNet** tiếng Anh [21].

Về phiên bản VietWordNet thì theo Virach Sornlertlamvanich 2010, vào thời điểm 05/12/2010, VietWordNet chứa 4,960 từ, 12,270 nghĩa (synset) và 15,872 cặp từ - nghĩa, ước tính hoàn thành 10.44% [16].

### Giới thiệu về kho ngữ liệu SentiWordNet và VietSentiWordNet

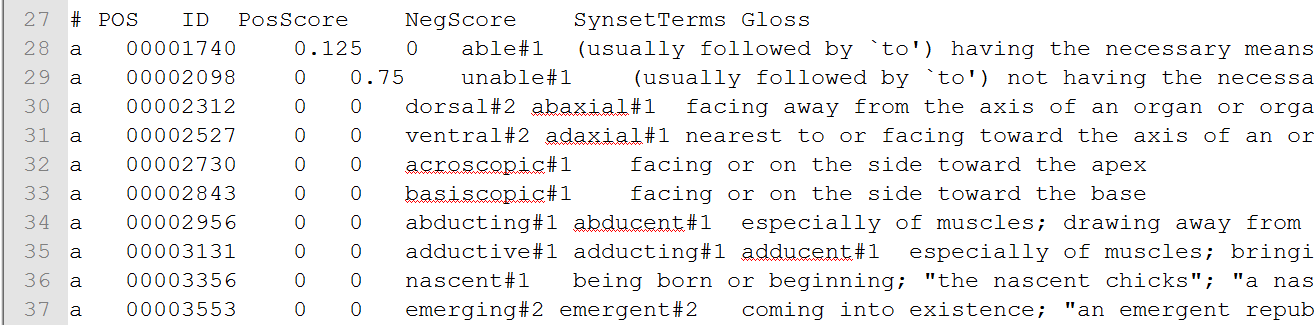
SentiWordNet là một cơ sở dữ liệu từ vựng được phát triển bởi Andrea Esuli và cộng sự, 2006 [3] dựa trên WordNet, nhằm hỗ trợ khai phá quan điểm.

Vũ Xuân Sơn và cộng sự, 2011 cũng đã xây dựng được bộ từ điển VietSentiWordNet ứng dụng khai phá quan điểm tin tức. Từ điển này có 977 synsets, với 1179 từ. Các tác giả áp dụng từ điển VietSentiWordNet vào bài toán phát hiện và tổng hợp quan điểm tin tức trên miền dữ liệu <http://vnexpress.net/> thu được độ chính xác tổng hợp quan điểm ở mức câu, cho kết quả cao nhất là 69,97% [1]. Trong khuôn khổ của khóa luận này, nhóm quyết định chọn bộ từ điển VietSentiWordNet cho quá trình tạo dữ liệu huấn luyện cho hệ thống để đưa ra quyết định phân lớp cảm xúc.

### Một số khái niệm được sử dụng về WordNet

* ***Synset:*** là một bản ghi, cấu tạo bởi 6 cột, các cột phân cách bởi dấu <tab>
  + - POS: từ loại của từ - part of speech
    - ID: mã đại diện cho synset
    - PosScore: trọng số tích cực của từ
    - NegScore: trọng số tiêu cực của từ
    - SynsetTerms: là những từ nhận định trong Synset.
* ***SynsetTerms:*** là những từ nhận định trong Synset. Một Synset có thể có chứa nhiều từ, và các từ này là đồng nghĩa với nhau. Một từ có thể có nhiều ngữ cảnh khác nhau và trọng số Pos(s)/Neg(s) sẽ khác nhau, do đó các từ này sẽ được gán kèm theo số hiệu để phân biệt các từ.
  + - ***Ví dụ:*** từ *high#3* có trọng số Pos(s)/Neg(s) là 0.375/0.125. Term *hope#1* có trọng số Pos(s)/Neg(s) là 0.125/0
* ***Gloss:*** là cột giải nghĩa và ngữ cảnh sử dụng của từ.

***Ví dụ:***



Hình 2.7: Cấu trúc trong file SentiWordNet 3.0

Trong đó:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *POS* | *ID* | *PosScore* | *NegScore* | *SynsetTerms* | *Gloss* |
| *a* | *00001740* | *0.125* | *0* | *able#1* | *(usually followed by to) having the necessary means* |
| *a* | *00002098* | *0* | *0.75* | *unable#1* | *(usually followed by to) not having the necessary means* |

Bảng 2.2: Mô tả cấu trúc của SentiWordNet 3.0

Phiên bản đầu tiên SentiWordNet 1.0 được xây dựng từ WordNet 2.0, mỗi synset được gán trọng số Obj(s) - độ khách quan, Pos(s) - độ tích cực , Neg(s) - độ tiêu cực của từ được chứa trong synset đó. Các nghĩa khác nhau của cùng một từ có thể có thuộc tính liên quan quan điểm (opinion-related) khác nhau [13].

Ví dụ: Trong SentiWordNet 1.0, synset [estimable (J,3)] tương ứng với nghĩa “*có thể đã tính toán hoặc ước lượng*” cho tính từ estimable, có điểm Obj(s) là 1.0, Pos(s) = Neg (s) =0, trong khi synset [estimable (J,1)] tương ứng với nghĩa “*xứng đáng được tôn trọng hoặc khen thưởng nhiều*” thì có Pos(s)=0.75, Neg(s)=0 và Obj(s) = 0.25.

Các trọng số Pos score, Neg score và Obj score được cho trong khoảng từ 0.0 đến 1.0, tổng ba trọng số này tƣng ứng với một synset là 1.0.

### Các phiên bản SentiWordNet

* SentiWordNet 1.0 được giới thiệu năm 2006 và được chấp nhận rộng rãi cho nghiên cứu. [3]
* SentiWordNet 1.1 được giới thiệu trong một bài báo cáo công nghệ năm 2007 nhưng không đạt tới mức công bố rộng rãi.
* SentiWordNet 2.0 được giới thiệu năm 2008 trong luận văn của Esuli [4].
* SentiWordNet 3.0 được giới thiệu năm 2010 [13].

SentiWordNet 1.x và 2.0 phát triển từ các chú thích trong WordNet 2.0. SentiWordNet 3.0 được phát triển dựa trên SentiWordNet 1.0 và WordNet 3.0.

### Giới thiệu bộ từ điển biểu tượng cảm xúc (emoticon)

Việc sử dụng các biểu tượng cảm xúc trên mạng xã hội là rất phổ biến. Các biểu tượng này thể hiện nhiều cảm xúc và rất có giá trị trong việc đánh giá cảm xúc của những bài đăng và bình luận. Vì vậy, việc xây dựng một bộ từ điển cho các biểu tượng cảm xúc là rất cần thiết. Khóa luận đã thực hiện xây dựng bộ từ điển Tiếng Việt cho các biểu tượng cảm xúc trên mạng xã hội Facebook dựa theo cấu trúc của SentiWorkNet để chuyển đổi những biểu tượng cảm xúc thành những từ vựng cảm xúc và sử dụng trong việc đánh giá cảm xúc của dữ liệu.

Bộ từ điển cảm xúc Tiếng Việt cho biểu tượng cảm xúc do nhóm xây dựng bao gồm 110 biểu tượng, bao phủ hầu như tất cả các biểu tượng cảm xúc được sử dụng trên mạng xã hội Facebook, tính đến ngày 25/12/2015.

## 

văn bản

đề

## Giới thiệu về phương pháp phân lớp SVM

Support Vector Machines (SVM) là một phương pháp phân loại xuất phát từ lý thuyết học thống kê, dựa trên nguyên tắc tối thiểu rủi ro cấu trúc (Structual Risk Minimisation). SVM cố gắng tìm cách phân loại dữ liệu sao cho lỗi xảy ra trên tập kiểm tra là nhỏ nhất (Test Error Minimisation). Đây là một phương pháp mới trong lĩnh vực trí tuệ nhân tạo. Vào thời kỳ đầu khi SVM xuất hiện, khả năng tính toán của máy tính còn hạn chế nên phương pháp SVM chưa được quan tâm. Từ năm 1995 trở lại đây, cùng với khả năng tính toán mạnh mẽ máy tính, các thuật toán sử dụng cho SVM phát triển rất nhanh và có nhiều ứng dụng to lớn.

### Ý tưởng của phương pháp

Cho trước một tập huấn luyện được biểu diễn trong không gian vector, trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu phẳng *f* quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp “+” và lớp “–”. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó khoảng cách biên càng lớn thì việc phân loại càng chính xác.

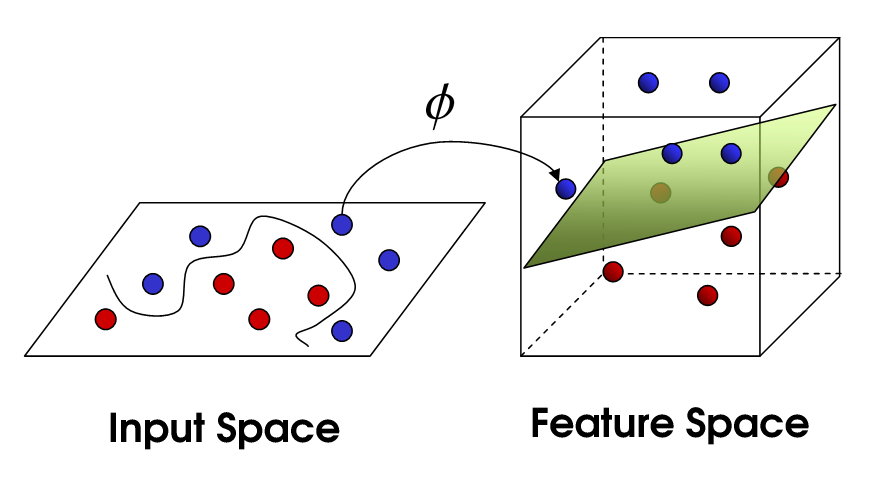
Mục đích của phương pháp SVM là tìm được một khoảng cách biên lớn nhất.

### Không gian đặc trưng

Sự hiệu quả của việc huấn luyện phụ thuộc nhiều vào cách dữ liệu được diễn tả. Khi dữ liệu được diễn tả phù hợp, việc huấn luyện sẽ dễ dàng và hiệu quả hơn. Vì vậy, việc chuyển đổi dữ liệu từ không gian đầu vào sang một không gian đặc trưng là rất cần thiết và là một bước quan trọng giúp dữ liệu được phân cụm tốt hơn. Ánh xạ vector:

Trong đó n là số chiều của đầu vào (hay còn gọi là số thuộc tính) và N là số chiều của không gian đặc trưng. Dữ liệu sẽ được chuyển vào không gian đặc trưng với N > n.

Không gian đặc trưng kí hiệu là F:



Hình 2.8: Ánh xạ vào không gian đặc trưng

### Cơ sở lý thuyết của phương pháp SVM

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

Cho tập mẫu với , thuộc vào hai lớp nhãn là nhãn lớp tương ứng của các

Phương trình siêu phẳng chứa vector trong không gian:

Đặt

Như vậy, biểu diễn sự phân lớp của vào hai lớp đã nêu.

### Các bước chính của phương pháp SVM

* Tiền xử lý dữ liệu: Xử lý các dữ liệu đầu vào nếu chúng chưa phải là số thực, đưa chúng về dạng số của SVM.
* Chọn hàm hạt nhân: Cần chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác phân lớp cao nhất.
* Thực hiện việc kiểm tra chéo để xác định các tham số phù hợp nhất
* Sử dụng các tham số cho việc huấn luyện tập mẫu
* Kiểm thử kết quả phân lớp với tập dữ liệu Test.

## Đánh giá độ chính xác của hệ thống

Đánh giá kết quả phương pháp phân lớp văn bản có thể được tính toán theo nhiều cách khác nhau. Theo khảo sát của Sebastiani [9], độ đo phổ biến nhất được sử dụng để đánh giá kết quả phân lớp là độ hồi tưởng và độ chính xác:

|  |  |  |  |
| --- | --- | --- | --- |
| **Lớp Ci** | | **Dữ liệu thực** | |
| Thuộc lớp Ci | Không thuộc lớp Ci |
| **Dự đoán** | Thuộc lớp Ci | **TPi** | **TNi** |
| Không thuộc lớp Ci | **FPi** | **FNi** |

Bảng 2.3: Các kí hiệu sử dụng trong đánh giá độ chính xác

Trong đó:

**TPi** (true positives): số lượng ví dụ positive được thuật toán phân đúng vào Ci

**TNi** (true negatives): số lượng ví dụ negative được thuật toán phân đúng vào Ci

**FPi** (false positives): số lượng ví dụ positive được thuật toán phân sai vào Ci

**FNi** (false negatives): số lượng ví dụ negative được thuật toán phân sai vào Ci

Độ chính xáccủa lớplà tỷ số ví dụ dương được thuật toán phân lớp, cho giá trị đúng trên tổng số ví dụ được thuật toán phân vào lớplà:

Độ hồi tưởngcủa lớplà tỷ lệ số lượng ví dụ dương được thuật toán phân lớp, cho giá trị đúng trên tổng số ví dụ dương thực sự thuộc lớplà:

Do nhóm thực hiện phân lớp tập dữ liệu thành 2 lớp, nên độ chính xác và độ hồi tưởng cần được tính cho toàn bộ tập 2 lớp. Nhóm sử dụng 3 độ đo là F1-Score, Precision và Recall để đánh giá độ chính xác của hệ thống với *D* là tập nhãn sẽ được phân lớp:

Precision:

Recall:

F1-Score:

## Giới thiệu về Apache Spark

### Tổng quan

Apache Spark là một open source cluster computing framework được phát triển sơ khởi vào năm 2009 bởi AMPLab tại đại học California, Berkeley. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay. Spark được xây dựng dựa trên tiêu chí đề cao tốc độ xử lý, dễ dàng sử dụng và có khả năng thực hiện các phân tích cực kì phức tạp.

Spark cho phép xây dựng và phân tích nhanh các mô hình dự đoán. Hơn nữa, nó còn cung cấp khả năng truy xuất toàn bộ dữ liệu cùng lúc. Thêm vào đó, Spark còn cung cấp tính năng *streaming,* được dùng để xây dựng các mô hình real-time bằng cách nạp toàn bộ dữ liệu vào bộ nhớ.

Khi ta có một tác vụ nào đó quá lớn mà không thể xử lý trên một laptop hoặc một server, Spark cho phép chúng ta phân chia tác vụ này thành những phần dễ quản lý hơn. Sau đó, Spark sẽ chạy các tác vụ này trong bộ nhớ, trên các cluser của nhiều server khác nhau để khai thác tốc độ truy xuất bộ nhớ của RAM. Spark sử dụng API Resilient Distributed Dataset (RDD) để xử lý dữ liệu.

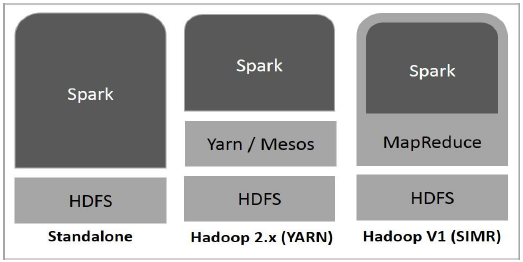
Spark nhận được nhiều sự hưởng ứng từ cộng đồng Big Data trên thế giới do cung cấp khả năng tính toán nhanh và nhiều thư viện đi kèm hữu ích như Spark SQL (với kiểu dữ liệu DataFrame), Spark Streaming, thư viện máy học (Mllib) hỗ trợ các chức năng chính: Phân lớp dữ liệu (Classification), hồi quy (Regression), gom nhóm (Clustering), các phương pháp lọc (Collaborative Filtering), giảm số chiều (Dimensionality Reduction), lựa chọn đặc trưng (Feature Extraction and Tranformations).

### Hadoop và Spark

Hadoop – một công nghệ xử lý dữ liệu BigData đã được khoảng 10 năm và được chứng minh là giải pháp lựa chọn để xử lý dữ liệu lớn. MapReduce là một giải pháp tuyệt vời cho việc tính toán theo một chiều, nhưng lại tỏ ra không hiệu quả cho tính toán cùng lúc và sử dụng các thuật toán. Mỗi bước trong một quá trính tính toán bao gồm một pha Map và một pha Reduce và bạn bắt buộc phải chuyển đổi sang mô hình này để tận dụng khả năng của MapReduce với tất cả các mục đích sử dụng.

Mỗi dữ liệu đầu ra giữa các bước của mỗi Job cần phải được lưu trữ trong hệ thống phân tán trước khi các bước tiếp theo có thể bắt đầu được. Do đó, phương pháp này có xu hướng chậm lại vì phải sao chép và lưu trữ dữ liệu trên ổ cứng. Ngoài ra, các giải pháp sử dụng Hadoop thường bao gồm các clusters, rất khó để thiết lập và quản lý. Nó cũng đòi hỏi người dùng phải tích hợp một số công cụ khác để có thể xử lý với các loại dữ liệu lớn khác nhau (như Mahout cho Machine Learning và Storm cho xử lý dữ liệu Streaming). Nếu người dùng muốn thực hiện những công việc phức tạp hơn, người dùng cần phải kết nối một loạt các thao tác MapReduce lại với nhau và thực hiện chúng theo tuần tự. Mỗi thành phần công việc lại có độ trễ cao, và các công việc không thể bắt đầu nếu như công việc ngay trước đó chưa kết thúc.

Với Spark các lập trình viên có thể thực hiện các công việc phức tạp cùng lúc với nhiều kiểu dữ liệu sử dụng mô hình đồ thị acyclic (DAG). Nó cũng hỗ trợ khả năng chia sẻ dữ liệu thông qua DAG, vì vậy các công việc khác nhau có thể thao tác với cùng một bộ dữ liệu chung. Spark hoạt động trên nên tảng của hệ thống file phân tán Hadoop Distributed File System (HDFS) cung cấp khả năng tăng cường và bổ sung các chức năng. Spark cũng hỗ trợ việc cài đặt trên một hệ thống Hadoop sẵn có (với SIMR – Spark – Inside – MapReduce) hoặc là phân cụm với Hadoop v2 YARN thậm chí là với Apache Mesos.



Hình 2.9: Ba cách để cài đặt Spark trên nền tảng Hadoop

Chúng ta nên nhìn nhận rằng Spark được thiết kế để thay thế cho Hadoop MapReduce hơn là một sự thay thế cho Hadoop. Nó hoàn toàn không có ý định thay thế Hadoop nhưng lại mang đến một giải pháp toàn diện và thống nhất để quản lý những yêu cầu và các trường hợp sử dụng BigData khác nhau.

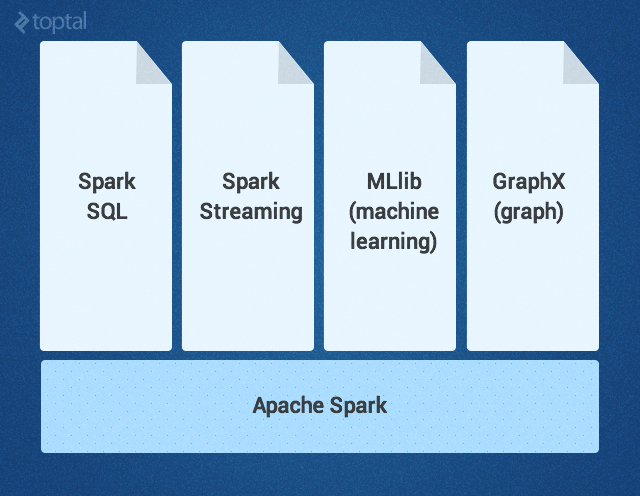
### Những tính năng nổi bật của Spark

* **Tốc độ:** Spark cho phép chương trình chạy trên một Hadoop cluster nhanh gấp 100 lần nếu chạy trên in-memory và hơn khoảng 10 lần nhanh khi chạy trên disk. Điều này có thể đạt được bằng cách giảm số lượng đọc/ghi vào đĩa. Nó lưu trữ dữ liệu vào các biến trung gian trong bộ nhớ.
* **Hỗ trợ cùng lúc nhiều ngôn ngữ lập trình:** Spark cung cấp các API được xây dựng sẵn cho Java, Scala và Python. Vì thế, bạn có thể tạo ra những ứng dụng bằng các ngôn ngữ lập trình khác nhau. Spark còn cung cấp hơn 80 function sẵn có cho các truy vấn tương tác.
* **Nâng cao khả năng phân tích:** Spark không chỉ cung cấp “Map” và “reduce”. Nó còn cung cấp những cơ chế hỗ trợ truy vấn SQL, Streaming data, Machine learning (ML), và các thuật toán Graph.

### Hệ sinh thái của Apache Spark

Bên cạnh các API của Spark Core, chúng ta sẽ có thêm rất nhiều các thư viện được tích hợp sẵn như là một phần của hệ sinh thái Spark và nâng cao khả năng trong việc phân tích dữ liệu BigData cũng như trong Machine Learnig.

Các thư viện đó bao gồm:



Hình 2.10: Hệ sinh thái Spark

* **Spark SQL:** là một thành phần trên cùng của Spark Core giới thiệu một khái niệm dữ liệu trừu tượng mới là SchemaRDD, cung cấp khả năng hỗ trợ cho các dữ liệu có cấu trúc hoặc bán – cấu trúc.
* **Spark Streaming:** cung cấp khả năng xử lý theo thời gian thực đối với dữ liệu streaming, chẳng hạn như các tài nguyên log file từ các webserver (ví dụ: dữ liệu từ Apache Flume and HDFS/S3), dữ liệu từ mạng xã hội như Twitter hay Facebook, và hàng đợi tin nhắn như Kafka.
* **MLlib:** là một thư viện Machine Learning cung cấp các thuật toán khác nhau được thiết kế để có thể thực hiện các bài toán thường gặp khi xử lý BigData như:
  + **Classification**: phân lớp bằng thống kê
  + **Regression**: phân tích hồi quy
  + **Collaborative Filtering:** các phương pháp lọc thông tin dùng cho hệ thống Recommendation
  + **Dimensionality Reduction:** quy trình rút gọn số biến ngẫu nhiên đang quan tâm, bao gồm lựa chọn biến và tạo biến mới
  + **Feature Extraction and Tranformations:** lựa chọn đặc trưng
* **GraphX:** là một thư viện dùng cho các thao tác đồ thị và thực hiện các hoạt động đồ thị song song. Nó cung cấp một công cụ đồng nhất cho ETL (Extract, Transform và Load), phân tích thăm dò và tính toán đồ thị. Ngoài xây dựng các thuật toán xung quanh các hoạt động về đồ thị, GraphX cũng cung cấp các giải pháp xử lý các thuật toán đồ thị phổ biến như PageRank.

**Kết luận chương hai:**

Chương hai của khóa luận trình bày về tất cả các kỹ thuật, cũng như tư tưởng chính để rồi đề xuất mô hình tổng hợp các “chủ đề” được thảo luận trên mạng xã hội Facebook dựa vào phương pháp LDA, sau đó thực hiện đánh giá cảm xúc theo từng chủ đề, cũng như sự kết hợp của từ điển lexicon VietSentiWordNet và phương pháp máy học SVM.

Trong chương tiếp theo, khóa luận sẽ tiến hành thực nghiệm trên phương pháp đã xây dựng, đánh giá kết quả đạt được của phương pháp và mô tả chi tiết của hệ thống dựa vào các màn hình giao diện hiển thị.

# XÂY DỰNG HỆ THỐNG TÌM KIẾM CHỦ ĐỀ ẨN VÀ ĐÁNH GIÁ CẢM XÚC TRÊN TRANG FACEBOOK

## Quy trình phát triển phần mềm

### Mô hình phát triển phần mềm

Ứng dụng được phát triển theo mô hình thác nước cải tiến bao gồm các giai đoạn sau:

* ***Phân tích và xác định các yêu cầu:*** Phân tích và xác định các yêu cầu của ứng dụng dựa trên nhu cầu nắm bắt thông tin của các cơ quan quản lý, giúp các cơ quan quản lý biết được những vấn đề mà được những thành viên trong một tổ chức quan tâm và phản ứng của họ về vấn đề đó. Từ đó, hỗ trợ trong việc ra quyết định cũng như các chính sách phù hợp.
* ***Thiết kế hệ thống và phần mềm:*** Thực hiện thiết kế hệ thống, sử dụng các thư viện, tìm hiểu các thuật toán, … để đáp ứng nhu cầu xử lý của hệ thống.
* ***Cài đặt hệ thống:*** Cài đặt các chức năng cho hệ thống như đã thiết kế, kiểm thử đơn vị cho từng chức năng (Unit test).
* ***Tích hợp và kiểm thử:*** Tích hợp các chức năng để triển khai hệ thống, chạy thử nghiệm và kiểm thử kết quả.
* ***Vận hành và bảo trì:*** Triển khai hệ thống thực tế và bảo trì, sửa lỗi hệ thống cũng như cải thiện độ chính xác cho hệ thống.

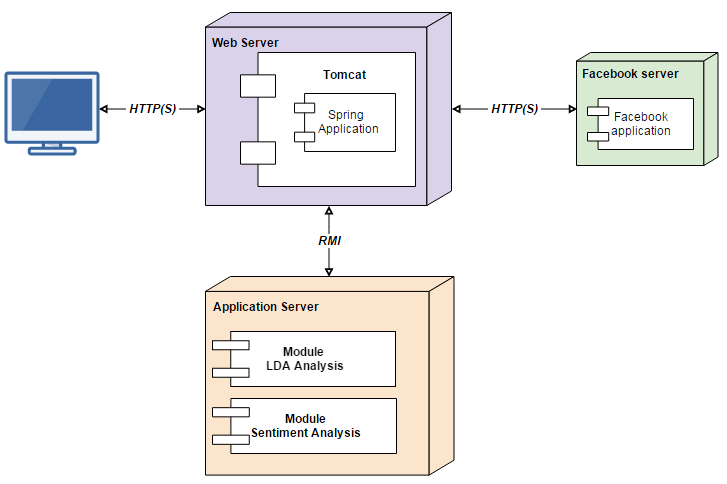


Hình 3.1: Mô hình thác nước cải tiến

### Xác định yêu cầu

Xây dựng ứng dụng web có chức năng khai thác dữ liệu từ Facebook và phân tích các dữ liệu đó để biết được các chủ đề chính mà một nhóm người cụ thể nào đó đang trao đổi. Sau khi biết nội dung các chủ đề đang được quan tâm, hệ thống phân tích và tiến hành phân lớp cảm xúc các nội dung theo từng chủ đề thành 3 lớp: Tích cực, tiêu cực và trung lập. Sau đó, hiển thị kết quả thống kê theo các lớp cảm xúc cho người dùng.

### Kiến trúc hệ thống



Hình 3.2: Kiến trúc hệ thống

***Mô tả***

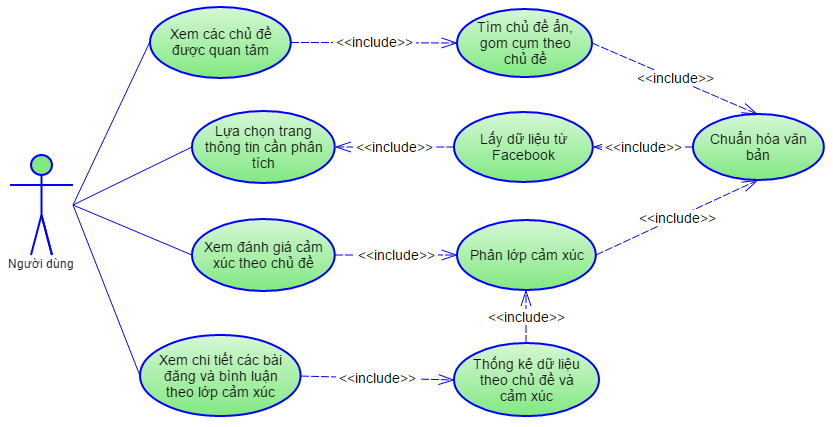
Hệ thống gồm 2 server chính là:

* Web Server: Cài đặt ứng dụng web dựa trên framework Spring. Server này sử dụng Apache Tomcat để triển khai ứng dụng web.
* Application Server: Cài đặt ứng dụng java để xử lý dữ liệu và các yêu cầu từ Web Server. Server này cài đặt Apache Spark để xử lý dữ liệu nhanh hơn và hỗ trợ quá trình huấn luyện cũng như quá trình trích xuất đặc trưng của dữ liệu đầu vào.

Hai server này kết nối với nhau bằng phương thức RMI (Remote Method Invocation) để việc truyền dữ liệu và gọi các hàm xử lý được đơn giản và dễ dàng hơn.

## Phân tích yêu cầu

### Sơ đồ use case

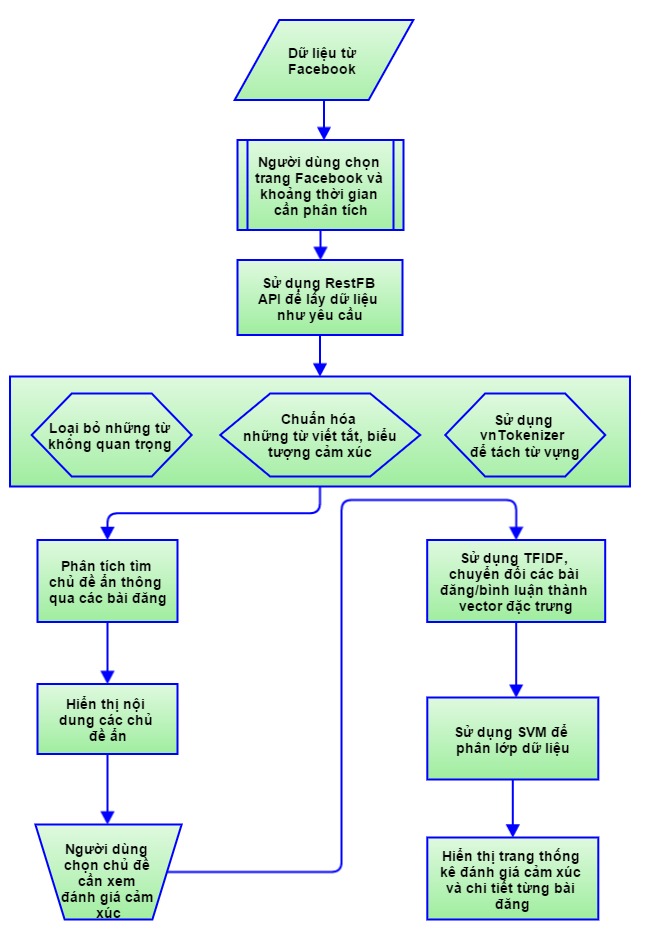


Hình 3.3: Lược đồ Use Case

### Mô tả sơ đồ use case

|  |  |  |
| --- | --- | --- |
| **Mã UC** | **Tên chức năng** | **Mô tả** |
| UC\_01 | Chọn trang thông tin cần phân tích và khoảng thời gian. | Đăng nhập Facebook với tài khoản của nhà phát triển để lấy được mã truy cập (Access token) hợp lệ.  Người dùng điền mã truy cập, chọn trang Facebook, chọn khoảng thời gian thích hợp |
| UC\_02 | Xem các chủ đề đang được quan tâm | Người dùng xem các chủ đề đang được trao đổi trên trang Facebook đã chọn. Mỗi chủ đề là một danh sách các từ vựng |
| UC\_03 | Xem đánh giá cảm xúc theo chủ đề | Chọn một chủ đề cụ thể để xem thống kê về đánh giá cảm xúc của hệ thống |
| UC\_04 | Xem chi tiết các bài đăng và bình luận | Chọn các bài đăng đã được đánh giá cảm xúc để xem chi tiết các cảm xúc của các bài đăng/bình luận |
| UC\_05 | Tìm chủ đề ẩn và gom cụm theo chủ đề | Sử dụng thuật toán LDA để tìm các chủ đề ẩn. Sử dụng WordCloud để hiển thị thông tin về các chủ đề. Phân cụm các bài đăng thành từng chủ đề |
| UC\_06 | Lấy thông tin từ Facebook theo yêu cầu | Sử dụng RestFB API để lấy dữ liệu công khai từ Facebook bao gồm: Những bài đăng, những bình luận thuộc bài đăng, ngày tháng của bình luận và bài đăng. Thông tin được lấy dựa theo các ID của các trang (fan-page) Facebook. |
| UC\_07 | Chuẩn hóa văn bản | Xây dựng bộ từ viết tắt, bộ từ vựng cho biểu tượng cảm xúc để chuẩn hóa cho dữ liệu lấy được từ Facebook. Loại bỏ những kí tự thừa, vô nghĩa để tạo được dữ liệu chuẩn cho bước tách từ. |
| UC\_08 | Phân lớp cảm xúc | Sử dụng bộ từ điển cảm xúc VietSentiWordNet để đánh giá cảm xúc cho những bài đăng/bình luận đã qua xử lý ban đầu, sau đó dùng kết quả để huấn luyện cho bộ phân lớp SVM. Sử dụng bộ phân lớp SVM để phân lớp cho những bài đăng/bình luận. |
| UC\_09 | Thống kê dữ liệu theo chủ đề và cảm xúc | Tạo giao diện web để người dùng tùy chọn những trang Facebook và khoảng thời gian cần phân tích. Hiển thị những chủ đề đang được trao đổi trong trang Facebook đó theo cụm các từ, sử dụng WordCloud để hiển thị. Khi người dùng chọn một chủ đề nào đó thì hệ thống sẽ hiển thị thống kê về số lượng/tỉ lệ các bài đăng/bình luận theo 3 lớp cảm xúc là: Tích cực, tiêu cực và trung lập, hệ thống cũng hiển thị chi tiết các bài đăng và bình luận liên quan đến bài đăng |

### Sơ đồ luồng dữ liệu toàn hệ thống



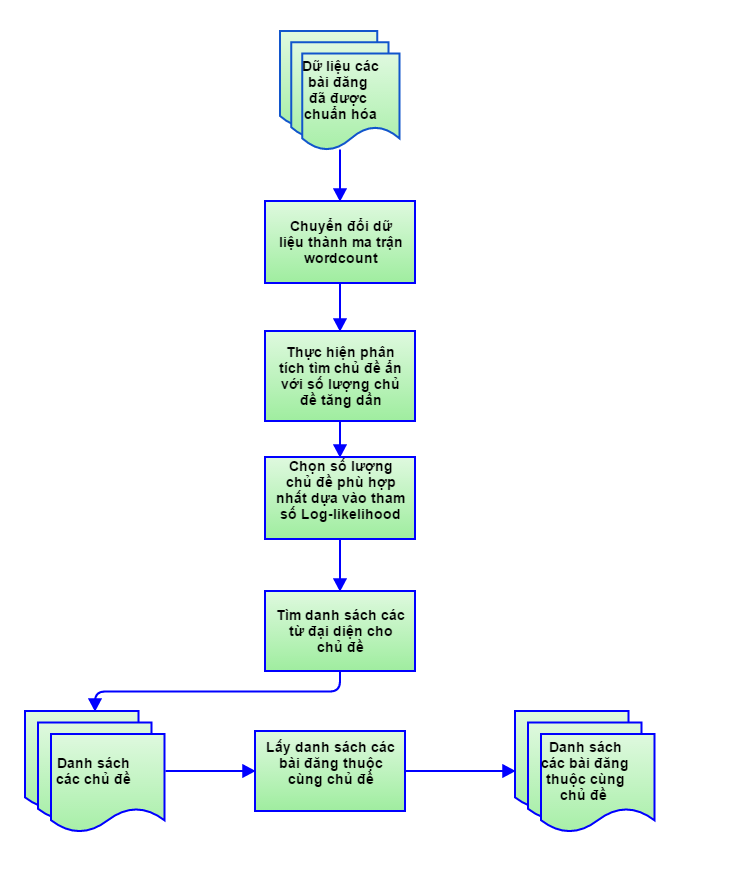
Hình 3.4: Sơ đồ DFD cho toàn hệ thống

**Mô tả sơ đồ:**

|  |  |
| --- | --- |
| **Bước** | **Mô tả** |
| 1 | Người dùng chọn trang và khoảng thời gian quan tâm để xem các chủ đề và đánh giá cảm xúc. |
| 2 | Hệ thống thực hiện lấy dữ liệu từ Facebook theo yêu cầu của người dùng thông qua RestFB API. |
| 3 | Xử lý ban đầu cho dữ liệu: Loại bỏ những kí tự không cần thiết, chuẩn hóa các kí tự viết tắt, chuyển đổi biểu tượng cảm xúc thành từ vựng, loại bỏ từ không quan trọng, không có nghĩa. Tách từ vựng Tiếng Việt |
| 4 | Hệ thống xử lý dữ liệu đã được chuẩn hóa và tách từ để tìm chủ đề ẩn, sử dụng thuật toán LDA. |
| 5 | Sau khi phân tích và tìm các chủ đề ẩn, hệ thống hiển thị thông tin về các chủ đề theo cụm các từ vựng lên giao diện người dùng |
| 6 | Người dùng xem nội dung các chủ đề, sau đó chọn một chủ đề bất kì để xem đánh giá cảm xúc về chủ đề đó. |
| 7 | Dựa vào những bài đăng và bình luận thuộc chủ đề được chọn, hệ thống sử dụng một thực thi của thuật toán TFIDF để thực hiện trích xuất những đặc trưng của từng tài liệu và chuyển đổi dữ liệu theo định dạng dữ liệu đầu vào của SVM. |
| 8 | Sử dụng bộ phân lớp SVM đã được huấn luyện để phân lớp cảm xúc cho các bài đăng và bình luận |
| 9 | Từ kết quả phân lớp, hệ thống thống kê số lượng và tỉ lệ các bài đăng/bình luận theo lớp cảm xúc (tích cực, tiêu cực, trung lập) và hiển thị kết quả thống kê lên giao diện người dùng. Hệ thống cũng hiển thị một cách trực quan nội dung các bài đăng và bình luận theo thứ tự các lớp cảm xúc. |

### Chức năng tìm chủ đề ẩn

#### Sơ đồ luồng dữ liệu

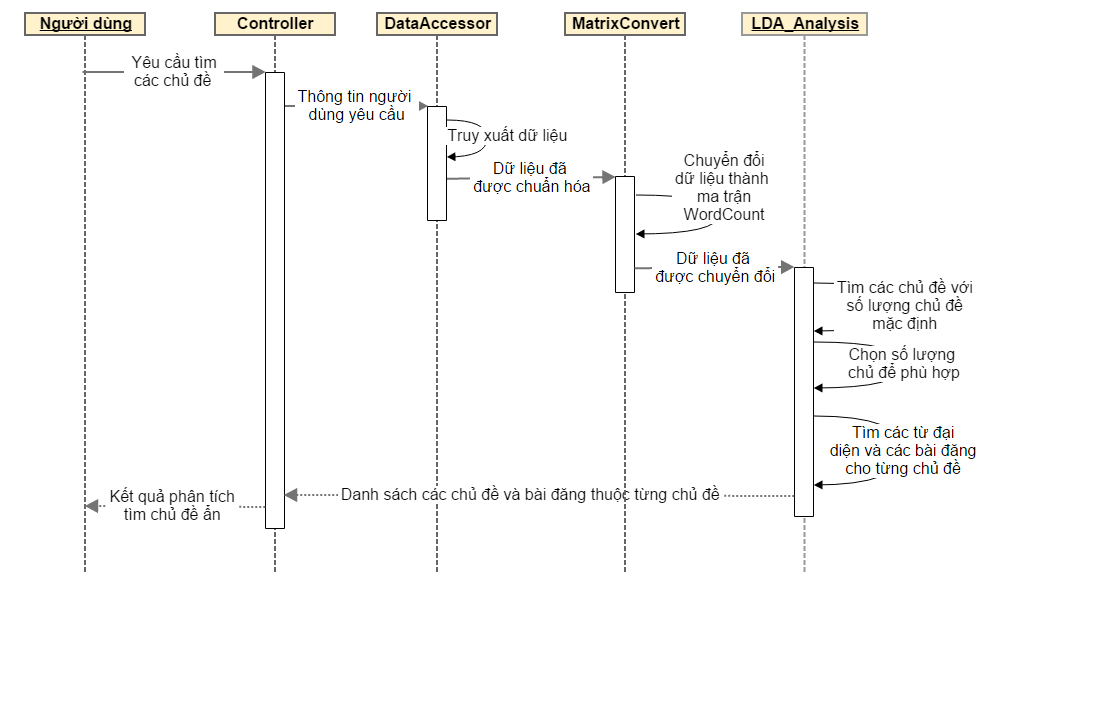


Hình 3.5: Sơ đồ DFD cho chức năng tìm chủ đề ần

**Mô tả sơ đồ:**

|  |  |
| --- | --- |
| **Bước** | **Mô tả** |
| 1 | Chuyển đổi dữ liệu thành ma trận word count (Số lần xuất hiện của mỗi từ là một giá trị trong ma trận) |
| 2 | Thực hiện phân tích chủ đề với số lượng chủ đề tăng dần đến số chủ đề mặc định |
| 3 | Tìm danh sách các từ đại diện cho chủ đề để tạo thành các chủ đề với số lượng từ mặc định, mỗi từ có một trọng số khác nhau |
| 4 | Tìm danh sách các bài đăng có cùng chủ đề để gom cụm theo chủ đề |

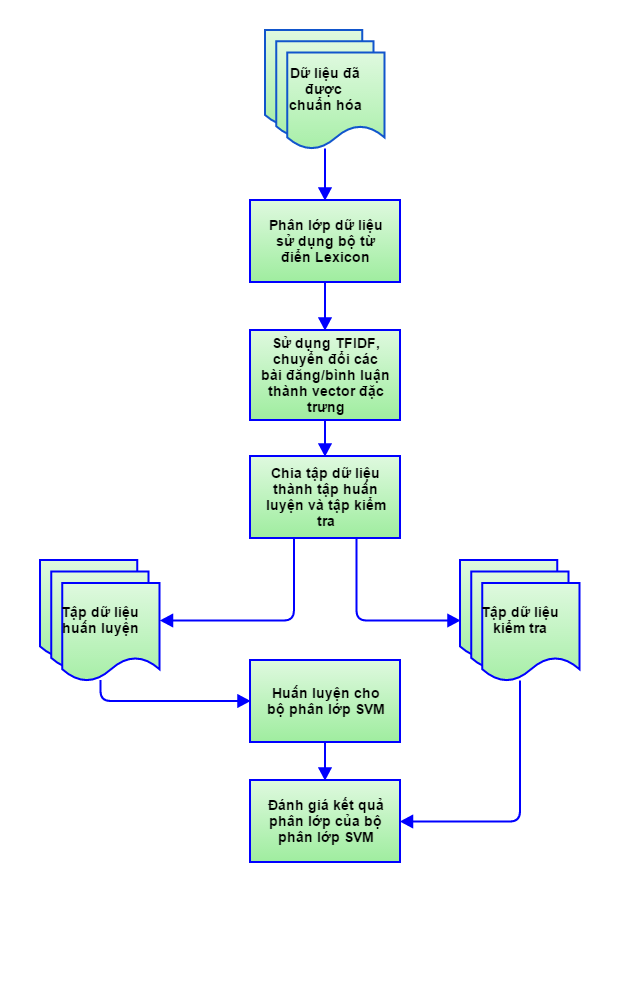
#### Sơ đồ Sequence



Hình 3.6: Sơ đồ Sequence cho chức năng tìm chủ đề ẩn

### Chức năng phân tích cảm xúc

#### Sơ đồ luồng dữ liệu



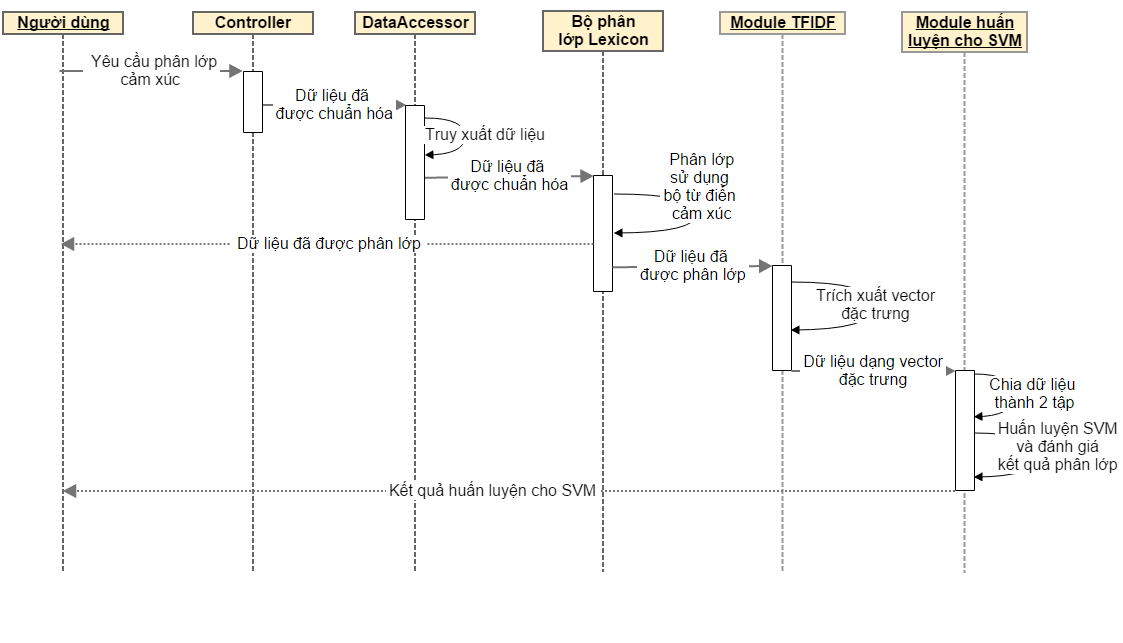
Hình 3.7: Sơ đồ DFD cho chức năng phân lớp cảm xúc

**Mô tả sơ đồ:**

|  |  |
| --- | --- |
| **Bước** | **Mô tả** |
| 1 | Phân lớp dữ liệu sử dụng bộ phân lớp Lexicon, thư viện từ điển cảm xúc VietSentiWordNet |
| 2 | Sử dụng một thực thi của thuật toán TFIDF để trích xuất đặc trưng của dữ liệu đã qua bước xử lý ban đầu và tạo định dạng dữ liệu để huấn luyện cho SVM |
| 3 | Chia bộ dữ liệu bao gồm các đặc trưng của dữ liệu thành hai tập riêng biệt một cách ngẫu nhiên là tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Tỉ lệ hai tập mà khóa luận sử dụng là 60% và 40%. |
| 4 | Sử dụng tập dữ liệu huấn luyện để huấn luyện cho bộ phân lớp SVM. |
| 5 | Sử dụng tập dữ liệu kiểm tra để đánh giá kết quả huấn luyện. |

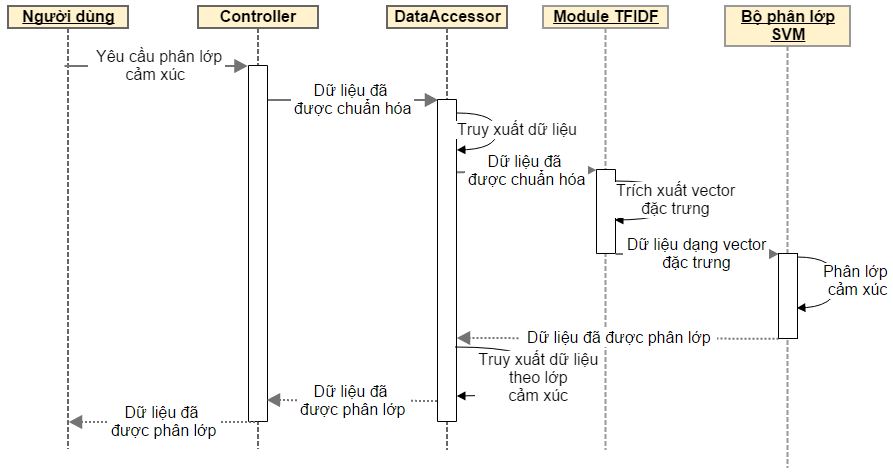
#### Sơ đồ Sequence

* **Giai đoạn huấn luyện cho bộ phân lớp SVM**



Hình 3.8: Sơ đồ Sequence giai đoạn huấn luyện cho SVM

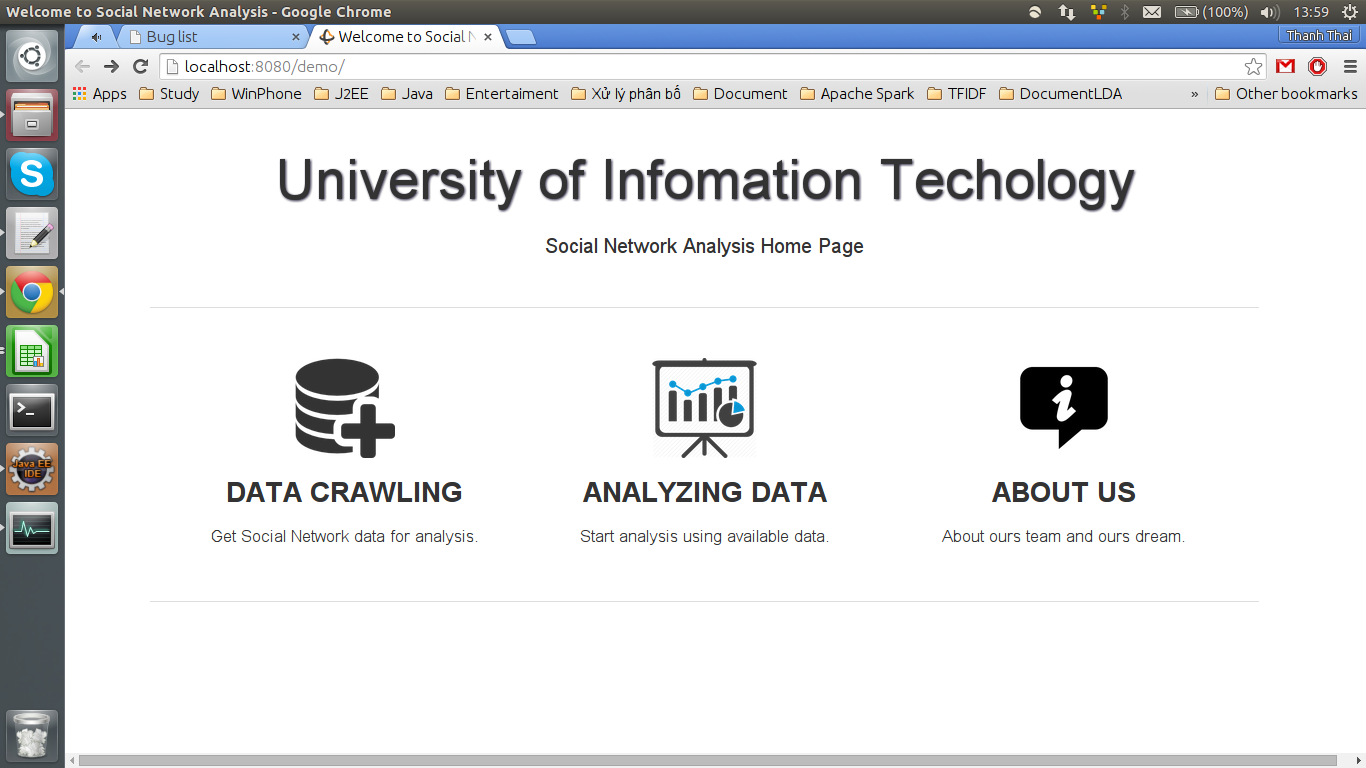
* **Giai đoạn phân lớp cảm xúc sử dụng SVM**



Hình 3.9: Sơ đồ Sequence giai đoạn phân lớp cảm xúc

## Thiết kế màn hình

### Màn hình chính

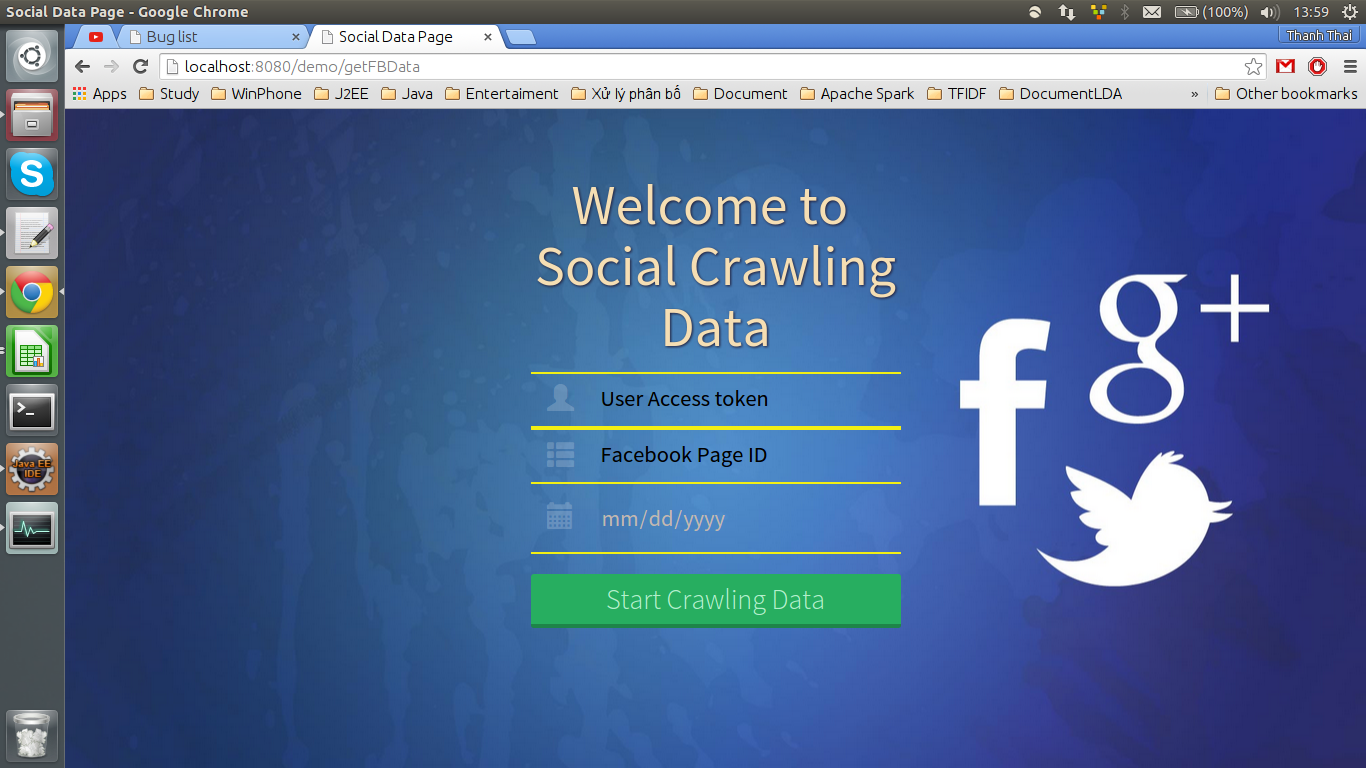


Hình 3.10: Màn hình chính

***Mô tả màn hình***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thành phần** | **Thể loại** | **Chức năng** |
| 1 | Data crawling | Button | Chọn để chuyển đến trang lấy dữ liệu từ Facebook |
| 2 | Data analyzing | Button | Chọn để sử dụng chức năng phân tích dữ liệu như: Tìm các chủ đề ẩn, đánh giá cảm xúc và xem thống kê |
| 3 | About us | Button | Chọn để xem thông tin về nhóm phát triển |

### Màn hình lấy dữ liệu từ Facebook

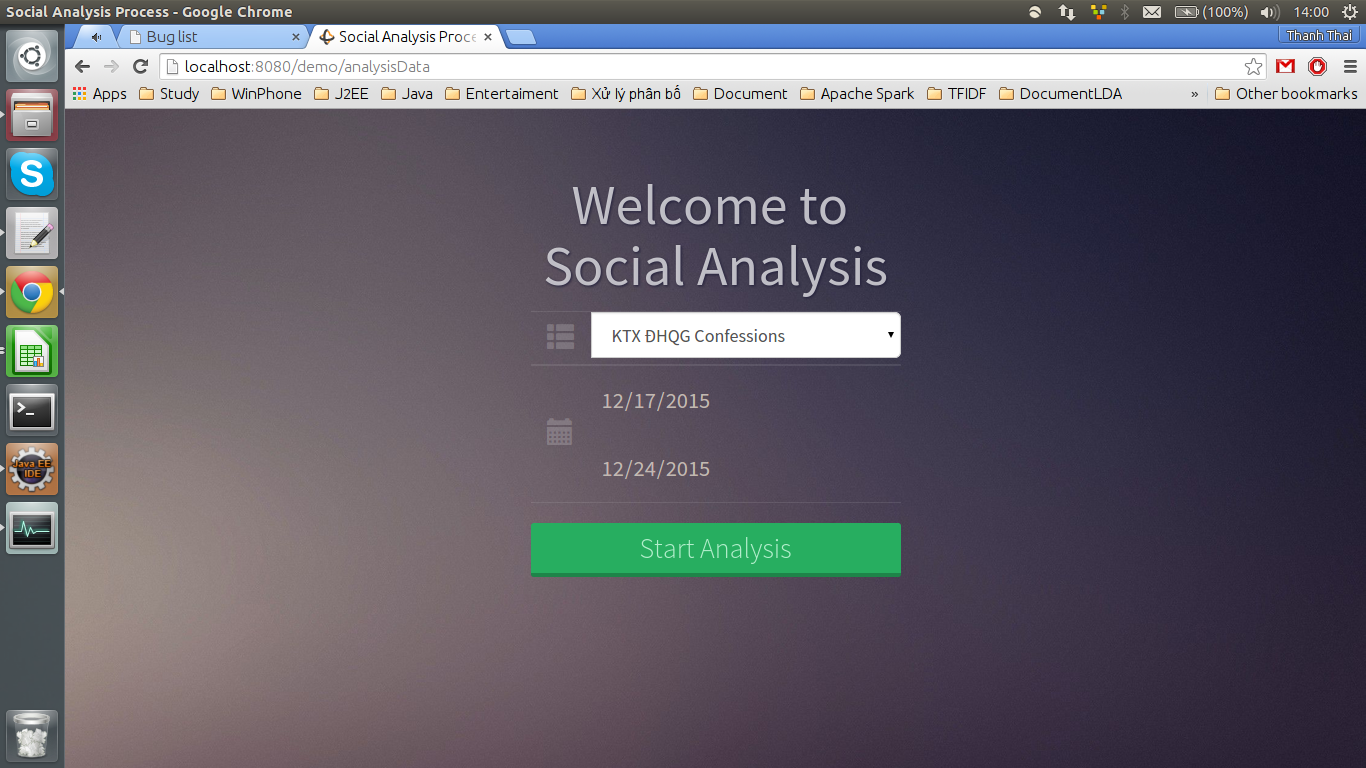


Hình 3.11: Màn hình lấy dữ liệu từ Facebook

***Mô tả màn hình***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thành phần** | **Thể loại** | **Chức năng** |
| 1 | User Access token | Textbox | Người dùng nhập mã truy cập (User Access token) của tài khoản nhà phát triển trên Facebook. |
| 2 | Facebook Page ID | Textbox | Người dùng nhập những mã của những trang Facebook muốn lấy dữ liệu |
| 3 | dtDateFrom | DateField | Người dùng nhập hoặc chọn ngày tháng bắt đầu lấy dữ liệu. Dữ liệu sẽ được lấy từ ngày này cho đến thời gian hiện tại |
| 4 | Start Crawling Data | Button | Chọn để bắt đầu quá trình lấy dữ liệu từ các trang Facebook và lưu vào cơ sở dữ liệu. |

### Màn hình chọn dữ liệu để phân tích

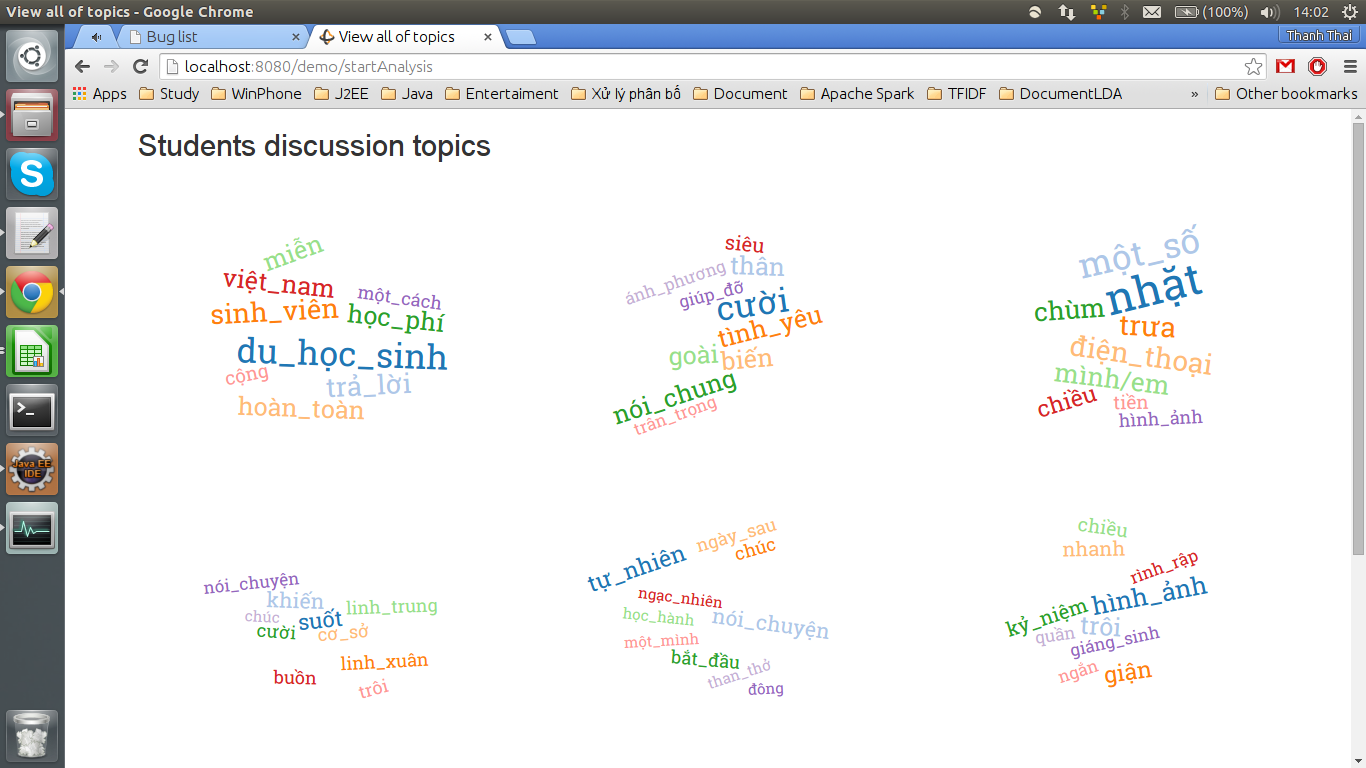


Hình 3.12: Màn hình chọn dữ liệu để phân tích

***Mô tả màn hình***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thành phần** | **Thể loại** | **Chức năng** |
| 1 | cboPageName | SelectBox | Người dùng chọn trang Facebook muốn phân tích từ danh sách. Danh sách trang bao gồm những trang mà người dùng đã crawl trước đó. |
| 2 | dtDateFrom | DateField | Người dùng nhập ngày bắt đầu của dữ liệu |
| 3 | dtDateTo | DateField | Người dùng nhập ngày kết thúc của dữ liệu |
| 4 | Start Analysis | Button | Chọn để hệ thống bắt đầu phân tích tìm các chủ đề đang trao đổi trên trang Facebook và trong khoảng thời gian đã chọn |

### Màn hình xem các chủ để đang trao đổi

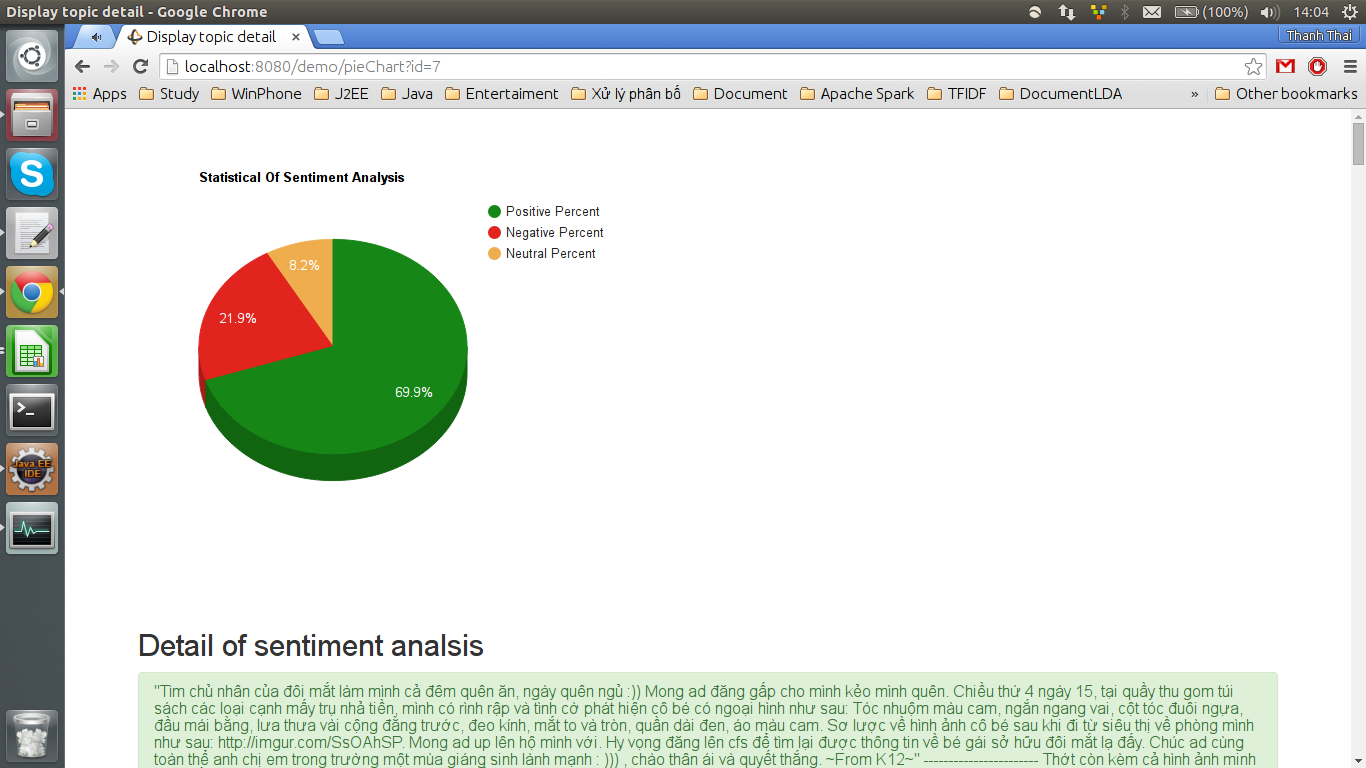


Hình 3.13: Màn hình xem các chủ đề đang được trao đổi

***Mô tả màn hình***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thành phần** | **Thể loại** | **Chức năng** |
| 1 | lstTopic\_n | List | Danh sách các chủ để, mỗi chủ đề gồm nhiều từ khác nhau. Người dùng chọn một chủ đề bất kì để xem đánh giá cảm xúc về chủ đề đó. |

### Màn hình xem thống kê về đánh giá cảm xúc và xem chi tiết bài đăng



Hình 3.14: Màn hình xem thống kê đánh giá cảm xúc

***Mô tả màn hình***

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thành phần** | **Thể loại** | **Chức năng** |
| 1 | ChartSumary | ChartCirle | Hiển thị thống kê % về các lớp cảm xúc (tích cực, tiêu cực, trung lập) của những bài đăng trong chủ đề được chọn |
| 2 | ClpPosts | CollapseLabel | Hiển thị chi tiết bài đăng với màu sắc tùy theo lớp. Khi người dùng chọn bài đăng thì hệ thống hiển thị những bình luận thuộc về bài đăng |
| 3 | lbComments | Label | Những bình luận thuộc về bài đăng được chọn. Những bình luận này cũng được hiển thị với màu sắc tương tự như các bài đăng. |

## Thực nghiệm và đánh giá

### Môi trường thực nghiệm

***Cấu hình phần cứng***

|  |  |
| --- | --- |
| **Thành phần** | **Thông số** |
| **CPU** | Intel Core i5 – 2.24 GHz |
| **RAM** | 6 GB |
| **Hệ điều hành** | MS Windows 8.1, 64-bit |
| **Bộ nhớ ngoài** | 500 GB |

Bảng 3.1: Cấu hình phần cứng cho thực nghiệm

***Công cụ và thư viện hỗ trợ***

|  |  |  |
| --- | --- | --- |
| **Tên phần mềm** | **Phiên bản** | **Nguồn** |
| **Eclipse IDE for Java** | Luna 3.5.7 | http://www.eclipse.org/downloads/ packages/eclipse-ide-java-ee-developers/lunar |
| **Java SDK** | 8.0 | http://www.oracle.com/technetwork/ java/javase/downloads/index.html |
| **vnTokenizer** | 4.1.1 | http://mim.hus.vnu.edu.vn/phuonglh/ softwares/vnTokenizer |
| **RestFB** | 1.18.0 | http://restfb.com/ |
| **VietSentiWorkNet** | 1.0 | http://sourceforge.net/projects /vietsentiwordne/ |
| **Apache Spark** | 1.5.2 | http://spark.apache.org/ |

Bảng 3.2: Công cụ và thư viện hỗ trợ

### Chuẩn bị dữ liệu

(TDL)đ nhưng chưa được.

Nhóm thực hiện 3 thực nghiệm trên 3 trang Facebook với 3 tập dữ liệu như sau:

***TDL 1: UIT Confessions***

* Địa chỉ trang Facebook: <https://www.fb.com/ConfessionUIT>
* Số lượng người dùng: 9 529 – ngày 25/12/2015
* Khoảng thời gian lấy dữ liệu: 01/09/2015 – 30/10/2015
* Tổng số bài đăng: 124
* Tổng số bình luận: 315

|  |  |
| --- | --- |
| **Tên tập** | **Số bản ghi** |
| Tập dữ liệu huấn luyện | 264 |
| Tập dữ liệu kiểm tra | 175 |
| Tổng | 439 |

***TDL 2: Nhân Văn Confessions***

* Địa chỉ trang Facebook: <https://www.fb.com/NhanVan.Cfs>
* Số lượng người dùng: 23 369 – ngày 25/12/2015
* Khoảng thời gian lấy dữ liệu: 01/09/2015 – 30/11/2015
* Tổng số bài đăng: 254
* Tổng số bình luận: 1037

|  |  |
| --- | --- |
| **Tên tập** | **Số bản ghi** |
| Tập dữ liệu huấn luyện | 775 |
| Tập dữ liệu kiểm tra | 516 |
| Tổng | 1291 |

***TDL 3: KTX ĐHQG Confessions***

* Địa chỉ trang Facebook: <https://www.fb.com/KTXDHQGConfessions>
* Số lượng người dùng: 31 844 – ngày 25/12/2015
* Khoảng thời gian lấy dữ liệu: 01/09/2015 – 30/10/2015
* Tổng số bài đăng: 514
* Tổng số bình luận: 2849

|  |  |
| --- | --- |
| **Tên tập** | **Số bản ghi** |
| Tập dữ liệu huấn luyện | 2018 |
| Tập dữ liệu kiểm tra | 1345 |
| Tổng | 3363 |

### Kết quả thực nghiệm

**Đánh giá kết quả phân cụm chủ đề bằng phương pháp LDA:**

Để đánh giá kết quả phân cụm chủ đề bằng LDA, khóa luận sử dụng độ đo Log-likelihood. Cũng theo Gregor Heinrich [10], đây là độ đo dùng để đánh giá kết quả ước lượng tham số cho mô hình, giá trị Log-likelihood càng cao, thì mô hình càng phù hợp.

Hình 3.15: Biểu đồ giá trị Log-likelihood theo số lượng chủ đề

Dữ liệu đánh giá được lấy từ fanpage UIT Confessions (<https://www.fb.com/ConfessionUIT>). Từ khoảng thời gian 01/12/2015 – 20/12/2015 với số lượng chủ đề tăng từ 2 đến 20. Kết quả thực nghiệm cho thấy giá trị số chủ đề bằng 9 là kết quả tối ưu nhất với giá trị

Log-likelihood = -12298.051010324742.

**Đánh giá kết quả phân tích cảm xúc kết hợp phương pháp lexicon và SVM:**

phân lớp cảm xúc

**Kết quả TDL 1:**

* Số lượng chủ đề: 5
* Độ chính xác phân lớp cảm xúc:

|  |  |  |  |
| --- | --- | --- | --- |
| **Độ đo** | **Tiêu cực** | **Tích cực** | **Trung bình hệ thống** |
| Precision | 0.62 | 0.81 | 0.76 |
| Recall | 0.71 | 0.78 | 0.76 |
| F1 score | 0.68 | 0.79 | 0.76 |

Hình 3.16: Đánh giá kết quả phân lớp với TDL1

**Kết quả TDL2:**

* Số lượng chủ đề: 7
* Độ chính xác phân lớp cảm xúc:

|  |  |  |  |
| --- | --- | --- | --- |
| **Độ đo** | **Tiêu cực** | **Tích cực** | **Trung bình hệ thống** |
| Precision | 0.7 | 0.65 | 0.68 |
| Recall | 0.62 | 0.74 | 0.68 |
| F1 score | 0.65 | 0.7 | 0.68 |

Hình 3.17: Đánh giá kết quả phân lớp với TDL2

**Kết quả TDL 3:**

* Số lượng chủ đề: 9
* Độ chính xác phân lớp cảm xúc:

|  |  |  |  |
| --- | --- | --- | --- |
| **Độ đo** | **Tiêu cực** | **Tích cực** | **Trung bình hệ thống** |
| Precision | 0.75 | 0.71 | 0.73 |
| Recall | 0.73 | 0.74 | 0.73 |
| F1 score | 0.74 | 0.72 | 0.73 |

Hình 3.18: Đánh giá kết quả phân lớp với TDL3

Dựa vào kết quả của 3 thực nghiệm trên đây cho thấy độ chính xác nhỏ nhất là: 65% và độ phủ nhỏ nhất là: 62%. Kết quả như vậy là khả quan và có thể áp dụng vào thực tế.

Tuy nhiên, 3 thực nghiệm trên đây chỉ mới sử dụng tập dữ liệu huấn luyện chưa qua kiểm tra và chỉnh sửa thủ công cho nên kết quả phân lớp cảm xúc chưa đạt độ chính xác cao do hạn chế của tập dữ liệu huấn luyện được phân lớp trước đó bằng cách sử dụng bộ từ điển VietSentiWordNet. Như đã đề cập ở phần trước, tập dữ liệu huấn luyện được chỉnh sửa thủ công sau khi đã phân lớp bằng bộ phân lớp Lexicon để đạt độ chính xác cao hơn. Nhóm đã tổng hợp dữ liệu của 3 thực nghiệm trên đây và chỉnh sửa lại nhãn phân lớp cho phù hợp. Sau đó tiến hành thực nghiệm nên tập dữ liệu này và cho kết quả như sau:

***TDL 4: Dữ liệu tổng hợp và đã chỉnh sửa nhãn phân lớp***

* Tổng số bài đăng: 1300
* Tổng số bình luận: 4000

|  |  |
| --- | --- |
| **Tên tập** | **Số bản ghi** |
| Tập dữ liệu huấn luyện | 3700 |
| Tập dữ liệu kiểm tra | 2300 |
| Tổng | 5000 |

**Kết quả TDL 4:**

* Số lượng chủ đề: 12
* Độ chính xác phân lớp cảm xúc:

|  |  |  |  |
| --- | --- | --- | --- |
| **Độ đo** | **Tiêu cực** | **Tích cực** | **Trung bình hệ thống** |
| Precision | 0.81 | 0.82 | 0.82 |
| Recall | 0.82 | 0.84 | 0.82 |
| F1 score | 0.81 | 0.83 | 0.82 |

Hình 3.19: Đánh giá kết quả phân lớp với TDL4

**Kết luận chương ba:**

Trong chương này, nhóm đã giới thiệu về chức năng cũng như thiết kế của hệ thống tìm chủ đề ẩn và phân tích cảm xúc các thành viên trên mạng xã hội Facebook. Qua kết quả của thiết kế cũng cho thấy hệ thống đáp ứng đủ yêu cầu về việc chọn dữ liệu cần phân tích, xem các chủ đề và xem đánh giá cảm xúc về các chủ đề đó. Thực nghiệm các chức năng cho kết quả khá tốt và khả quan để có thể áp dụng vào thực tế

# KẾT LUẬN

## Kết quả đạt được

Qua quá trình tìm hiểu về bài toán phân tích trạng thái cảm xúc, khóa luận đã thấy được vai trò của cảm xúc, và khả năng ứng dụng của việc phân tích trạng thái cảm xúc vào việc đưa ra quyết định dựa trên các kỹ thuật đánh giá cảm xúc. Khóa luận cũng đã thực hiện được việc phát hiện các chủ đề đang được quan tâm nhất và thống kê mức độ đánh giá/cảm xúc của các thành viên trên mạng xã hội Facebook với từng chủ đề ấy.

Trong khuôn khổ các nghiên cứu phục vụ quá trình hoàn thiện khóa luận, nhóm đã đạt được những kết quả cụ thể sau:

* Tìm hiểu và xây dựng module hỗ trợ rút trích dữ liệu từ facebook, sử dụng thư việc RestFb và các thao tác với Graph API.
* Tìm hiểu các phương án gom cụm dữ liệu phổ biến, các đặc điểm chính cũng như ưu điểm và khuyết điểm của từng phương pháp. Sau đó nhóm đi sâu nghiên cứu và chọn LDA là giải pháp gom cụm dữ liệu vì những ưu điểm và khả năng làm việc với dữ liệu lớn.
* Hiểu cách xây dựng và sử dụng được thư viện từ vựng cảm xúc – VietSentiWordNet.
* Tạo ra bộ từ điển về các biểu tượng cảm xúc (emoticons) Tiếng Việt đối với mạng xã hội Facebook, nâng cao khả năng phân tích các giá trị cảm xúc.
* Đề xuất và cài đặt thử nghiệm về việc mô hình kết hợp phân tích cảm xúc bằng Lexicon (dùng từ điển VietSentiWordNet) sau đó kết hợp với phương pháp SVM để phân lớp từng cảm xúc cụ thể. Kết quả cài đặt và thực nghiệm cho thấy, mô hình này có độ chính xác khá cao cũng như có độ phủ lớn, có khả năng ứng dụng vào thực tế.
* Nghiên cứu và ứng dụng các phương thức Machine Learning được cài đặt và hỗ trợ bởi Apache Spark (<http://spark.apache.org/>) giúp tăng tốc độ của hệ thống cũng như khả năng mở rộng để giải quyết bài toán trong trường hợp dữ liệu lớn.
* Ngoài ra, việc xây dựng hệ thống thử nghiệm đã giúp cho nhóm:
* Kiểm nghiệm được những kiến thức lý thuyết và ý tưởng để giải quyết bài toán làm thế nào phân cụm được cảm xúc.
* Có cái nhìn rõ ràng hơn về việc đặt ra vấn đề, tìm hiểu các kỹ thuật, phương pháp, và ứng dụng nó để giải quyết các vấn đề ban đầu.
* Bên cạnh đó, nâng cao khả năng tìm hiểu, phân tích, nghiên cứu các kỹ thuật, cũng như thiết kế hoàn thiện hệ thống. Nhằm giải quyết các bài toán, các vấn đề.

## Hướng phát triển

Trong thời gian tới, ngoài việc tiếp tục giải quyết các vấn đề còn tồn tại, nhóm sẽ định hướng một số nghiên cứu tiếp theo:

* Thực hiện thực nghiệm trên các miền dữ liệu khác, bao gồm các mạng xã hội khác cũng như các trang báo điện tử, các trang chuyên về đánh giá các sản phẩm.
* Thực hiện thực nghiệm với các phương pháp lựa chọn đặc trưng khác, mở rộng khả năng phân lớp dữ liệu cảm xúc bằng các phương pháp máy học.
* Cập nhật lại các giá trị chưa chính xác và thêm vào một số từ mới cho bộ từ điển VietSentiWordNet đã có.
* Xây dựng hoàn thiện bộ từ điển chuyên dùng cho việc đánh giá cảm xúc cho các mạng xã hội.
* Thực hiện việc đánh giá ở mức quan điểm, tức là quan điểm của chủ thể, đối với từng vấn đề riêng biệt.

**TÀI LIỆU THAM KHẢO**

**TÀI LIỆU TIẾNG VIỆT**

[1] Vũ Xuân Sơn, Trần Trung Hiếu, Lê Thu Hà, Đào Thủy Ngân. Xây dựng từ điển VietSentiWordNet ứng dụng khai phá quan điểm trên tin tức. *CÔNG TRÌNH THAM GIA GIẢI THƯỞNG “SINH VIÊN NGHIÊN CỨU KHOA HỌC” NĂM 2011, ĐẠI HỌC CÔNG NGHỆ.*

**TÀI LIỆU TIẾNG ANH**

[2] Alessia D'Andrea and Fernando Ferri and Patrizia Grifoni and Tiziana Guzzo, *“Approaches, Tools and Applications for Sentiment Analysis Implementation”,* September 2015. Published by Foundation of Computer Science (FCS), NY, USA. BibTeX

[3] Andrea Esuli, Fabrizio Sebastiani (2006) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC’ 6* *Bing Liu (2010).), enova, IT*

[4] Andrea Esuli. 2008. Automatic Generation of Lexical Resources for Opinion Mining: Model, Algorithms, and Application. Ph.D. thesis Scuola di Dottorato in Ingegneria Leonardo da Vinci, University of Pisa, Pisa, IT.

[5] Bing Liu, *Web data mining; Exploring hyperlinks, contents, and usage data, 2006, chapter 11: Opinion Mining*. Springer.

[6] Bing Liu (2010). Sentiment Analysis and Subjectivity. *Invited Chapter for the and book of Natural Language Processing, Second Edition. March, 2010*

[7] Bo Pang and Lillian Lee, “*Opinion Mining and Sentiment Analysis*”, Foundations and Trends R in Information Retrieval, 2, 1–2, 2008, pp. 1–135.

[8] D. Blei, A., Ng, and M. Jordan. “*Latent Dirichlet Allocation*”. In Journal of Machine Learning Research, 2003.

[9] Fabrizio Sebastiani. “*Machine Learning in Automated Text Categorization*”.

ACM Computing Survey, 34(1) pages 1-47, 2002.

[10] Gregor Heinrich, Technical Report Fraunhofer IGD Darmstadt, Germany: *“Parameter estimation for text analysis”,* 15 September 2009.

[11] Jasmeet Kaur and Neha Singh, “*Facebook Integration with RESTFB API*”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014.

[12] Kushal Dave, Steve Lawrence, and DavidM. Pennock, “*Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*”, In Proceedings of WWW, 2003, pp. 519–528.

[13] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani (2010). *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, *LREC 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010*

[14] Tausczik, Y.R. & Pennebaker, J.W. 2010. *The psychological meaning of words: Liwc and computerized text analysis methods*. Journal of Language and Social Psychology.

[15] Thomas Hofmann, “*Probabilistic latent semantic indexing*”, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999

[16] Virach Sornlertlamvanich (2010). *AsianWordNet: What’s next? DD6&AWN2010, December 7-9, 2010 , Phuket, Thailand*

[17] William M. Darling, “*A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling*”, December 1, 2011.

[18] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558. Copyright © 2010 (American Society for Information Science and Technology)

**WEBSITE TIẾNG VIỆT**

[19] Thuật toán tách từ Tiếng Việt: <http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/thuat-toan-tach-tu-tokenizer/thuat-toan-tach-tu>

**WEBSITE TIẾNG ANH**

[20] WordNet 3.0: <https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

[21] WordNet tiếng Anh: <http://wordnet.princeton.edu/>

[22] A hybrid approach to word segmentation of Vietnamese texts: http://mim.hus.vnu.edu.vn/phuonglh/node/33