**3. Khai thác dữ liệu mạng xã hội Twitter**

**Giới thiệu**

Sự bùng nổ ngày càng mạnh mẽ của mạng xã hội mở ra nhiều cơ hội cho các tổ chức, cá nhân thu thập, tìm kiếm thông tin cũng như xử lý chúng và nhiều bài toán được đặt ra để khai thác nguồn thông tin dồi dào từ mạng xã hội.

Khai thác dữ liệu là một trong các bài toán khai thác thông tin nằm trong ứng dụng của xử lý ngôn ngữ tự nhiên, giúp thu thập được những thông tin mong muốn về quan điểm của dữ liệu. Bài toán phân tích các đánh giá cho một chủ đề nhất định, hoặc sự kiện, sản phẩm để tự động phân loại đánh giá theo hướng tích cực, tiêu cực của quan điểm. Bài toán còn bao gồm nội dung tổng hợp quan điểm từ các tài liệu quan điểm thu được

Với sự phát triển nhanh chóng và mạnh mẽ, mạng xã hội Twitter đã trở thành một nguồn cung cấp nhiều thông tin quan điểm cho những người nghiên cứu về lĩnh vực này. Đã có nhiều công trình nghiên cứu liên quan đến bài toán khai phá quan điểm theo các phương pháp khác nhau từ nhiều nguồn dữ liệu khác nhau, nhưng Twitter là một nguồn dữ liệu phổ biến.

Đánh giá về quan điểm của bài viết trong Twitter vào một trong hai lớp Tích cực hoặc Tiêu cực mang lại nhiều ý nghĩa cho nhiều lĩnh vực như kinh tế, quảng cáo v.v... Một trong những cách tiếp cận là sử dụng học máy thống kê. Trong cách tiếp cận này, biểu diễn đặc trưng đóng vai trò quan trọng, ảnh hưởng trực tiếp tới độ chính xác của bộ phân lớp. Thông qua tìm hiểu, phân tích các phương pháp khai phá quan điểm trên tập dữ liệu Twitter, chúng tôi tập trung tới ba phương pháp biểu diễn đặc trưng: N-gram[7], Độ đặc trưng dựa trên tâm (Center-base similarity – CBS) [6] và Logcount ratio[9] để thực hiện đề tài chuyên mục: “Khai thác dữ liệu mạng xã hội Twitter”.

Nội dung gồm :

**3.1 Giới thiệu về khai thác dữ liệu mạng xã hội**

**3.2 Tổng quan về khai thác dữ liệu mạng xã hội Twitter**

**3.3 Một số phương pháp khai thác dữ liệu**

**3.4 Phân loại ý kiến khi khai khác trên Twitter**

**3.5 Kết quả dữ liệu thực nghiệm và đánh giá**

**3.1 Giới thiệu về khác thác dữ liệu mạng xã hội**

Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính..

Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.

Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến.

Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

**3.1.1 Một vài ví dụ trong khai thác dữ liệu**

Bài toán xác định quan điểm

Bài toán phân loại quan điểm

Bài toán khai phá quan điểm theo đặc trưng, khía cạnh

Bài toán tóm tắt quan điểm

**3.1.2 Một số khó khăn trong quá trình khai thác dữ liệu**

Phong cách viết của mỗi người là khác nhau

Sự phụ thuộc vào bối cảnh và thời gian

Tồn tại nhiều quan điểm mâu thuẫn trong cùng một tài liệu

**3.2 Tổng quan về khai thác dữ liệu mạng xã hội Twitter**

**3.2.1 Mạng xã hội Twitter là gì?**

Twitter là một dịch vụ mạng xã hội trực tuyến miễn phí cho phép người sử dụng đọc, nhắn và cập nhật các mẩu tin nhỏ gọi là tweets, một dạng tiểu blog.

Những mẩu tweet được giới hạn tối đa 280 ký tự được lan truyền nhanh chóng trong phạm vi nhóm bạn của người nhắn hoặc có thể được trưng rộng rãi cho mọi người.

**3.2.2 Tại sao phải chọn Twitter?**

Mạng xã hội Twitter ngày càng lớn mạnh, cùng với đó, việc thể hiện quan điểm trên mạng xã hội này cũng càng phổ biến. Việc khai thác dữ liệu trên mạng xã hội này có nhiều ý nghĩa, giúp ta thấy quan điểm của một người về một vấn đề nào đó.

**3.2.3 khó khăn và thách thức khi khai thác dữ liệu**

Việc phân loại quan điểm với dữ liệu Twitter gặp khá nhiều khó khăn và thách thức.

Các tweet có đặc điểm riêng khác so với các dạng văn bản truyền thống hay dữ liệu trên các forum, mạng xã hội khác như Facebook.

Số ký tự tối đa nhỏ, từ viết tắt, từ lóng, từ sai chính tả, hơn nữa, chất lượng và độ tin cậy thấp.

Nhiều đặc điểm của dữ 4 liệu Twitter làm giảm hiệu quả khai phá quan điểm dựa trên những kỹ thuật xử lý ngôn ngữ tự nhiên truyền thống.

**3.3 Một số phương pháp khai thác dữ liệu**

Các phương pháp ***cây quyết định***

* + - **Cây hồi quy** (*Regression tree*) ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)
    - **Cây phân loại** (*Classification tree*), nếu *y* là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

Phương pháp ***K-Nearest Neighbor***

* + - K-Nearest Neighbor được mô tả như sau:
    - Xác định giá trị tham số K (số láng giềng gần nhất)
    - Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng các Euclidean)
    - Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point
    - Lấy tất cả các lớp của K láng giềng gần nhất đã xác định
    - Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point

Thuật toán ***SVM (Support Vector Machine)***

là một khái niệm trong [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) và [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) cho một tập hợp các phương pháp [học có giám sát](https://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t) liên quan đến nhau để [phân loại](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_lo%E1%BA%A1i_(h%E1%BB%8Dc_m%C3%A1y)&action=edit&redlink=1) và [phân tích hồi quy](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_quy).

SVM gồm 3 dạng:

1. SVM tuyến tính
2. SVM dạng ban đầu
3. SVM dạng đối ngẫu

Ngoài ra còn 1 số phương pháp khác…

Một số nghiên cứu liên quan:

* + - Đã có rất nhiều công trình nghiên cứu của các tác giả liên quan đến việc khai phá quan điểm từ nhiều nguồn dữ liệu với các phương pháp khác nhau ví dụ như :
      * Agarwal, Apoorv and Xie, Boyi and Vovsha, Ilia and Rambow, Owen and Passonneau, Rebecca (2011), Sentiment Analysis of Twitter. Columbia University. New York, NY 10027 USA.
      * Davidov, Dmitry and Tsur, Oren and Rappoport, Ari (2010). Enhanced sentiment learning using Twitter hashtags and smileys, in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 241-249.

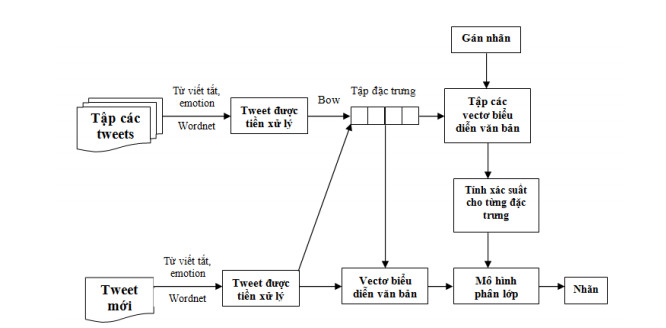
**3.4 Phân loại ý kiến khi khai khác trên Twitter**

Phân loại ý kiến trên Twitter là phân loại cho từng bình luận theo hướng quan điểm tích cực hay tiêu cực dựa trên nội dung bình luận. Trong bài viết này, chúng tôi đề xuất sử dụng mô hình túi từ và giải thuật máy học Multinomial Naïve Bayes để phân loại ý kiến.

Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, chúng tôi tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa, biểu diễn văn bản theo mô hình túi từ.

Giai đoạn tiền xử lý cho ra tập dữ liệu có số chiều lớn, nhưng trong đó đa số (khoảng 99%) các giá trị bằng 0. Để tiết kiệm bộ nhớ, chiến lược lưu trữ chỉ lưu những giá trị khác 0 (theo định dạng LibSVM). Cách lưu trữ này dẫn đến yêu cầu cài đặt lại giải thuật máy học Multinomial Naïve Bayes để có thể xử lý định dạng mới của tập dữ liệu.

Kết quả thực nghiệm trên các tập dữ liệu cho thấy bản cài đặt mới của giải thuật Multinomial Naïve Bayes (MNB) phân lớp hiệu quả, đơn giản và chính xác.



Hình: Sơ đồ phân lớp ý kiến với giải thuật MNB

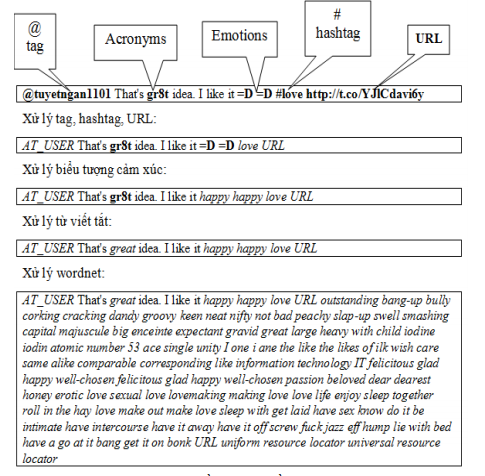
**3.4.1 Phương pháp tiền xử lý dữ liệu**

Dữ liệu của mạng Twitter rất phức tạp, phi cấu trúc, nhiễu. Để có thể phân lớp bằng giải thuật máy học, trước hết cần phải thực hiện các thao tác tiền xử lý.

Các tên tài khoản Twitter của người dùng được chuyển về dạng USER\_AT.

Địa chỉ trang web được chuyển về dạng URL.

Chúng tôi còn xử lý thêm các ký tự trùng lắp gần nhau, biểu tượng cảm xúc, từ viết tắt, tiếng lóng, mạng ngữ nghĩa.

 Hính: ví dụ về các bước tiền xử lý

**3.4.2 Biểu diễn dữ liệu**

Mô hình Bag of Words (BoW) là một mô hình được sử dụng phổ biến trong lĩnh vực phân loại văn bản. Mô hình này thường sử dụng để xử lý ngôn ngữ tự nhiên, được dùng để biểu diễn tài liệu, xem tài liệu là một tập hợp các từ (words) mà không quan tâm đến thứ tự cũng như cấu trúc cú pháp của chúng.

Giả sử dữ liệu có 15.000 tweets với 20.000 đặc trưng (từ vựng), thông thường mỗi tweet sẽ được lưu trữ như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chỉ mục | 1 | 2 | 3 | ... | 20.000 |
| Tần số  xuất hiện | 0 | 1 | 0 | … | 0 |

Đối với nghiên cứu này, chúng tôi đề xuất cách lưu trữ tiết kiệm bộ nhớ, tương tự như LibSVM [Chang & Lin, 2011], chỉ lưu những từ có tần số xuất hiện lớn hơn 0. Cách lưu trữ như sau:

**<label><index-1>:<value-1><index-2>:<value-2>…**

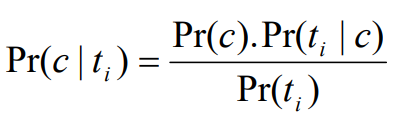
Trong đó:

* + - **<Lable>**là lớp ban đầu của tweet, 1 là tích cực, 0 là tiêu cực.
    - <**index-i>** chỉ mục của từ thứ i.
    - < **value-i**> tần số xuất hiện của từ i.

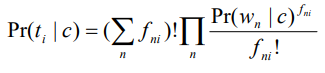
**3.4.3 phân loại ý kiến bằng giải thuật máy học**

Multinomial Naïve Bayes (MNB) là một mô hình đơn giản nhưng hoạt động rất tốt trong việc phân loại văn bản. [Lewis & Gale, 1994] đã đề xuất kết hợp mô hình túi từ và NB tạo ra giải thuật Multinomial Naïve Bayes. Cụ thể trong bài toán của chúng ta như sau:

Gọi C là tập hợp các lớp của văn bản (C có 2 phần tử +1 và -1). Gọi ti là một văn bản mới đến. Ta chọn xác suất để ti thuộc vào lớp ci lớn nhất. Xác suất này được tính bởi công thức:



* + Với c ∈ C
  + Pr(c) được tính bằng tổng số văn bản của lớp c chia cho tổng số văn bản của tất cả các lớp.
  + Khi tìm giá trị lớn nhất của Pr(c|ti) ta có thể bỏ qua tính Pr(ti) do không đổi khi so sánh.
  + Xác suất Pr(ti|c) được tính bằng công thức:



* fni là tần suất từ thứ n trong ti
* Pr(wn|c) là xác suất của từ thứ n khi cho trước lớp c.

**3.5 Kết quả dữ liệu thực nghiệm và đánh giá**

**3.5.1 Dữ liệu**

Để chuẩn bị dữ liệu cho việc thực nghiệm, chúng tôi xây dựng một module để tiến hành thu thập dữ liệu từ Twitter thông qua Twitter’s Streaming API.

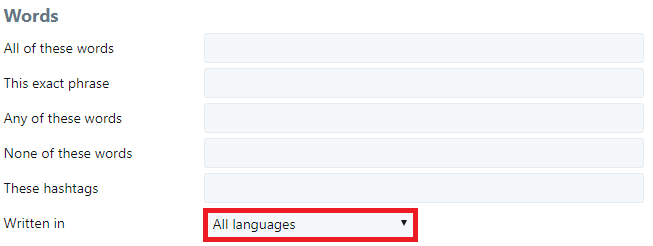
Sau khi thu thập dữ liệu từ Twitter, chúng tôi tiến hành chuẩn hóa dữ liệu bằng 4 thao tác sau trước khi sử dụng cho thực nghiệm:

Link thực hiện 3 thao tác (1)(2)(3): [**https://twitter.com/search-advanced**](https://twitter.com/search-advanced)

Link tham khảo Apache Tika toolkit:

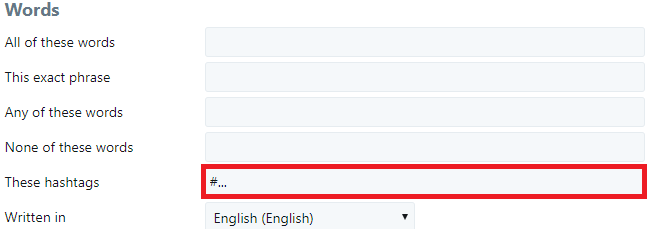
<https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html>

1. lọc bỏ những tweets mà không thuộc ngôn ngữ tiếng Anh và không có chứa URL

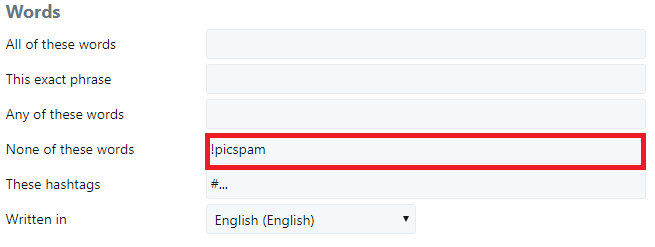
****

Ở phần All languages ta chọn English.

1. trích xuất những chú thích (hashtag) và địa chỉ web (URL) từ tweets



1. loại bỏ những chú thích vô nghĩa như “!picspam”, “atthissummer” dựa trên từ điển WordNet

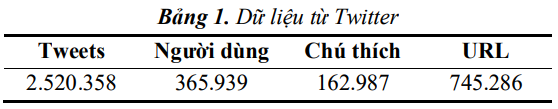


1. loại bỏ những địa chỉ trang web mà không thuộc ngôn ngữ tiếng Anh thông qua Apache Tika toolkit.

Bảng Tóm tắt Phương pháp Apache Tika toolkit.

|  |  |
| --- | --- |
| Công cụ sửa đổi và loại | Phương thức và mô tả |
| static void | [**addProfile**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#addProfile-java.lang.String-org.apache.tika.language.LanguageProfile-)([**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) language, [**LanguageProfile**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageProfile.html) profile)  Thêm 1 hồ sơ ngôn ngữ duy nhất |
| static void | [**clearProfiles**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#clearProfiles--)()  Xóa hồ sơ của ngôn ngữ hiện tại |
| static [**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) | [**getErrors**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#getErrors--)()  Trả về chuổi thông báo liên quan đến khởi tạo ngôn ngữ |
| [**String**](http://docs.oracle.com/javase/7/docs/api/java/lang/String.html?is-external=true) | [**getLanguage**](https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html#getLanguage--)()  Nhận ngôn ngữ được xác định |
| … | Chi tiết thao khảo thêm ở Link sau: <https://tika.apache.org/1.18/api/org/apache/tika/language/LanguageIdentifier.html> |

Bảng 1 dưới đây mô tả số liệu cụ thể của dữ liệu thu thập sau khi đã được chuẩn hóa:



**3.5.2 Phương pháp đánh giá**

**Sơ lược về MAP**

|  |  |
| --- | --- |
| MAP là gì? | MAP chỉ là mức trung bình của AP, hoặc độ chính xác trung bình, cho tất cả người dùng.  Ví dụ: Nếu chúng ta có 1000 người dùng, chúng tôi tính tổng số AP cho mỗi người dùng và chia tổng số cho 1000. Đây là MAP. |
| AP là gì? | Chúng tôi có thể đề xuất tối đa x mục cho mỗi người dùng  AP trả tiền để gửi tất cả các khuyến nghị x, bởi vì chúng tôi không bị phạt vì những dự đoán xấu  Các vấn đề trật tự, do đó, tốt hơn nên gửi thêm một số đề xuất nhất định, tiếp theo là các đề xuất chúng tôi ít chắc chắn hơn  Vì vậy, về cơ bản chúng tôi chọn x ứng cử viên tốt nhất (theo thứ tự) và đó là nó AP |
| Tài liệu tham khảo thêm | Nếu bạn cần hiểu thêm về Map có thể truy cấp vào đường link sau: **http://fastml.com/what-you-wanted-to-know-about-mean-average-precision**/ |

**Đánh giá thực nghiệm**

Ví dụ: Nếu chúng ta có 1000 người dùng, chúng tôi tính tổng số AP cho mỗi người dùng và chia tổng số cho 1000. Đây là MAP.

Ý tưởng của việc đánh giá được thực hiện như sau: Cho câu truy vấn q = {t} được nhập bởi người dùng u với từ khóa truy vấn t, kết quả tìm kiếm liên quan là những trang web được người dùng u với chú thích bằng từ khóa t.

Chúng tôi sử dụng độ đo Mean Average Precision để tiến hành thực nghiệm cho công cụ tìm kiếm. Cụ thể hơn, chúng tôi tính toán MAP cho mỗi người dùng và sau đó tính toán giá trị trung bình của tất cả các giá trị MAP (Mean MAP).



Trong đó, MAPi là giá trị độ đo MAP của người dung thứ i, Nu số lượng người dùng trong thực nghiệm.

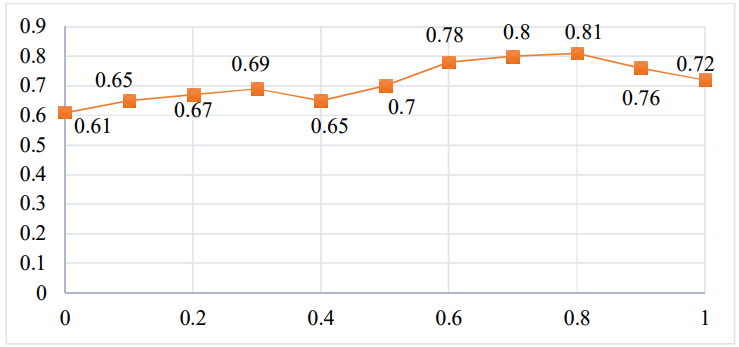
Chúng tôi chọn ngẫu nhiên 1000 cặp (u,t). Trong mỗi cặp, user u gửi câu truy vẫn q = {t} đến hệ thống tìm kiếm. Tiếp theo, hệ thống sẽ tìm kiếm và xếp hạng kết quả tìm kiếm phù hợp với câu truy vấn.

Cuối cùng, chúng tôi tính độ đo MAP cho 1000 câu truy vấn. Đánh giá một liên kết trả về có tương quan hay không tương quan tùy vào ý kiến chủ quan của người dùng, trong thực nghiệm chọn 10 kết quả đầu tiên để đánh giá.

**3.5.3 Kết quả đánh giá**

Trong công thức xếp hạng trang web (2), chúng tôi thực nghiệm với 0 ≤ α ≤ 1 và β = 0.5.

Công thức (4), chúng tôi chọn γ = 0.5 và kích thước câu truy vấn cho mỗi từ khóa là 5. Hình 1 dưới đây mô tả kết quả thực nghiệm của hệ thống tìm kiếm.



Giá trị α

**Hình 1.** Giá trị Mean Average Precision theo α

Theo như hình 1, hiệu quả của hệ thống tìm kiếm đạt tốt nhất khi α ∈ [0.6, 0.8].