Đánh giá sản phẩm điện tử dựa trên nhận xét của người dùng trên internet.

Lê Văn Duyệt, Trần Võ Tân Nguyên, TS. Ngô Thanh Hùng
Trường Đại học Công nghệ Thông tin – ĐHQG-HCM
{13520171, 14520613}@gm.uit.edu.vn, hungnt@uit.edu.vn

Tóm tắt: *Trong bài báo này, nhóm đề tài sử dụng bộ công cụ rút trích để thu thập thông tin phản hồi/bình luận từ các trang diễn đàn công nghệ và trang thương mại điện tử lớn. Sử dụng hai kỹ thuật máy học là Naïve Bayes và SVM để phân lớp ý kiến bình luận. Kết quả của phân lớp sẽ được đánh giá để tìm ra phương pháp tối ưu hơn, và kết quả sẽ được phục vụ cho mục đích tạo báo cáo, biểu đồ.*

Từ khóa: khai phá quan điểm, phân loại văn bản, rút trích, SVM, Naïve Bayes

1 Giới thiệu

Hầu hết các doanh nghiệp đều luôn muốn quan tâm đến ý kiến, phản hồi của khách hàng về sản phẩm/dịch vụ của họ như thế nào. Các đánh giá của khách hàng một mặt giúp cho những người dùng khác định hướng trong việc chọn lựa sản phẩm, mặt khác giúp cho các doanh nghiệp định hướng cải tiến chất lượng. Số lượng đánh giá về một sản phẩm mà chúng ta nhận được ngày càng tăng và có thể đến từ nhiều nguồn khác nhau (web bán hàng, diễn đàn, blog, mạng xã hội, ...). Vì vậy, để có tổng hợp ý kiến phản hồi chất lượng, thì phải tự động hóa được công việc thu thập và phân tích đánh giá.

Phân lớp văn bản là bài toán cơ bản trong khai phá quan điểm. Các hệ thống phân lớp văn bản là các hệ thống phải có khả năng xác định, khai phá ra nội dung thông tin. Có thể coi phân lớp quan điểm là bài toán phân lớp văn bản theo hai lớp tích cực và tiêu cực. Do đó một số kỹ thuật phân lớp văn bản như K-means, Naïve Bayes, Maximum entropy và SVM có thể sử dụng trong phương pháp học máy phân lớp quan điểm.

2 Tình hình nghiên cứu

Một đánh giá của khách hàng thường được thể hiện dưới dạng văn nói; một đánh giá thường chứa nhiều lỗi chính tả, chữ viết tắt, thành ngữ [11] và có thể được viết bởi giới trẻ với các ký hiệu, ngôn ngữ teen, pha trộn ngôn ngữ cũng như các biểu tượng cảm xúc [7]. Việc phân tích đánh giá của một sản phẩm theo thời gian, từng phiên bản của sản phẩm cũng là một vấn đề khó khăn.

Ngoài ra bài toán phân tích cũng phụ thuộc rất nhiều vào thang đánh giá. Thang đánh giá là phân loại đánh giá của khách hàng. Hầu hết các hệ thống đều phân loại "Negative", "Possitive", có một số trường hợp phân tích "Very bad", "Bad", "Satisfactory", "Good", "Excellent"; các đánh giá này có thể sắp xếp từ 1-5 sao. Bên cạnh đó cũng có hệ thống sử dụng "Thumbs up" và "Thumps down" [12].

Nhìn chung, có hai hướng tiếp cận chính đã được các nhà nghiên cứu sử dụng trong phân tích đánh giá: sử dụng máy học và dựa vào từ điển. Ngoài ra cũng có một số hệ thống phân tích đánh giá sử dụng kết hợp cả hai hướng tiếp cận trên [7].

Trong hướng tiếp cận máy học cần hai tập dữ liệu: Tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Các thuật toán máy học đã được các nhà nghiên cứu sử dụng thành công, mang lại kết quả cao là Support Vector Machine (SVM), Naïve Bayes (NB) và Maximum Entropy (ME). Thuật toán máy học cần xác định đặc trưng và các đặc trưng thường được sử dụng:

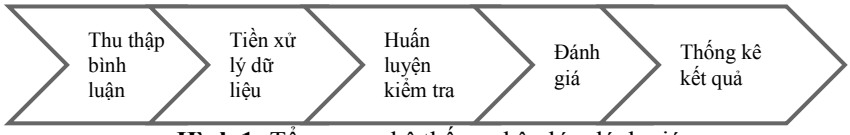
- Tần số xuất hiện: Với mô hình n-gram và uni-gram được xem là đặc trưng, phương pháp này đã được nghiên cứu nhiều và đạt kết quả tốt. B. Pang và nhóm tác giả [7] sử dụng mô hình uni-gram trên các đánh giá về phim của khách hàng, K. Dave và nhóm tác giả [4] sử dụng bi-grams và tri-grams trên các đánh giá về sản phẩm;
- Part of speech (POS): Các từ trong đánh giá được gắn thẻ POS để sử dụng làm đặc trưng. R. Prabowo và M. Thelwall [13] sử dụng POS trong nghiên cứu của họ tạo một tập các đặc trưng với tính từ và trạng từ. M. Hu và B. Liu [13] sử dụng thẻ POS và một số kỹ thuật xử lý ngôn ngữ tự nhiên để lấy tính từ và các từ cảm xúc;
- Phủ định: B. Pang và L. Lee [3] sử dụng để tìm các từ trái nghĩa, là một yếu tố quan trọng trong phân tích đánh giá. B. Pang và nhóm tác giả [7] đã sử dụng ba thuật toán máy học: Support Vector Machine, Naïve Bayes và Maximum Entropy. Họ đã so sánh hiệu quả và đi đến một kết luận rằng, Naïve Bayes thực hiện tốt trên tập đặc trưng nhỏ, Support Vector Machine thực hiện tốt trên không gian đặc trưng lớn, Maximum Entropy cho kết quả tốt hơn so với Naïve Bayes khi thử nghiệm với không gian đặc trưng lớn [7].

Trong các thuật toán máy học, SVM được chứng minh là công cụ phân lớp mạnh, hiệu quả hơn phân lớp văn bản truyền thống như Naïve Bayes. Thêm vào đó, B.Pang và các cộng sự [4] áp dụng kỹ thuật Naïve Bayes, maximum entropy và SVM để xác định hướng quan điểm phân cực trong bình luận về phim. Kết quả phân lớp sử dụng mô hình unigram và phân lớp SVM đạt kết quả cao nhất 82.9%. Điều đó cho ta thấy rằng SVM vẫn là một công cụ hiệu quả cho phân lớp quan điểm.

Phân lớp tài liệu theo hướng quan điểm thật sự là vấn đề thách thức và khó khăn trong lĩnh vực xử lý ngôn ngữ đó chính là bản chất phức tạp của ngôn ngữ của con người, đặc biệt là sự đa nghĩa và nhập nhằng nghĩa của ngôn ngữ. Sự nhập nhằng này rõ ràng sẽ ảnh hưởng đến độ chính xác bộ phân lớp của chúng ta một mức độ nhất định. Ví dụ câu sau: *"Làm thế nào để ai đó có thể ngồi xem hết bộ phim này?"* không chứa ý có nghĩa duy nhất mà rõ ràng là nghĩa tiêu cực. Theo đó, quan điểm dường như đòi hỏi sự hiểu biết nhiều hơn, tinh tế hơn.

3 Mô hình phân lớp đánh giá sản phẩm

Hệ thống nhóm đề tài triển khai gồm các module chính được thực hiện tuần tự theo các bước sau:



Hình 1: Tổng quan hệ thống phân lớp đánh giá

3.1. Rút trích bình luận

Module rút trích nhóm phát triển trên công nghệ Nodejs[1] và cơ sở dữ liệu lưu trữ MongoDB[2]. Cơ sở dữ liệu MongoDB có vai trò lưu thông tin các liên kết cần phải duyệt qua và kết quả thu thập được. Khởi đầu, crawler sẽ được cung cấp một số liên kết (URL) khởi đầu và tiến hành thu thập toàn bộ nội dung HTML chứa thông tin. Mã HTML được phân tích cấu trúc DOM[3], theo các luật quy định sẵn, crawler sẽ xác định vùng dữ liệu cần bóc tách: liên kết tương tự, thông tin bình luận cần thu thập. Các liên kết được chọn lọc và lưu trữ trong một hàng đợi URL. Để rút trích đúng mục tiêu bài viết, các liên kết hoặc tiêu đề bài viết được lọc lại theo từ khóa ứng với sản phẩm cần thu thập. Quá trình này đi lặp lại, cho tới khi không còn liên kết nào trong hàng đợi hoặc đủ số lượng cần thiết.

3.2. Tiền xử lý dữ liệu

Dữ liệu sau khi rút trích được tiền xử lý để có được một tập dữ liệu rõ ràng, không trùng lấp, loại bỏ các liên kết, trích dẫn trong bình luận.

Từng dữ liệu sẽ được tách từ, đối với tiếng Anh, các từ được phân cách bởi dấu cách hoặc các dấu câu khác. Tuy nhiên, một từ tiếng Việt có thể gồm nhiều hơn một âm tiết. Nhóm sử dụng thư viện tách từ vnTokenizer[4] của tác giả Lê Hồng Phương [9] với độ chính xác 96% đến 98%, lọc bỏ stopword bằng cách sử dụng từ điển stopword tiếng Việt[5] kết hợp loại bỏ theo tần suất xuất hiện của từ (chỉ số TF*IDF). Những từ có giá trị TF*IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

Kế đến, mỗi văn bản di trong tập ngữ liệu đang xét sẽ được mô hình hóa như là một vector trọng số của các đặc trưng, $d_i(w_{i1},\ldots,w_{im})$. Trong bài viết này, trọng số của một

[1] https://nodejs.org
[2] https://www.mongodb.org
[3] Document Object Model (DOM) - http://www.w3.org/DOM/
[4] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
[5] https://github.com/duyetdev/vietnamese-stopwords

từ được tính theo tần suất xuất hiện của từ trong văn bản (TF) và tần suất nghịch đảo của từ trong tập ngữ liệu (IDF).

$$w_{ij} = TF_{ij} \times IDF_{ij} = TF_{ij} \times \log\left(\frac{N}{DF_j}\right)$$

- TF_{ij} là số lần xuất hiện của từ thứ j trong văn bản thứ i.
- DF_j là tổng số văn bản có chứa từ thứ j trong tập ngữ liệu.
- N là tổng số văn bản trong tập ngữ liệu.

Ví dụ trong: D = *"Xiaomi cấu hình khủng mà giá thì bình dân. Giá này rất hợp với túi tiền"*.

Sau khi tách từ và lọc stopword ta được: `d1 = {"xiaomi", "cấu hình", "khủng", "giá", "bình dân"}, d2 = {"giá", "hợp", "túi tiền"}.`

Ví dụ với "Xiaomi": `tf("xiaomi", d1) = 1; idf("xiaomi") = log(2/1) = 0.3; tf*idf = 0.3`

Với "giá": `tf("giá", d1) = 1; idf("giá") = log(2/2) = 0; td*idf = 0`

Mô hình hóa, ta được vector đặc trưng:

`d1(0.3, 0.3, 0.3, 0, 0); d2(0, 0.3, 0.3)`

3.3. Phân lớp phản hồi, bình luận

Phản hồi sau khi được thu thập sẽ được phân thành các lớp khác nhau để phục vụ mục việc thống kê, tạo báo cáo.

Phân lớp văn bản (Text Classification) là quá trình gán nhãn các văn bản ngôn ngữ tự nhiên một cách tự động vào môt hoặc nhiều lớp cho trước, "nhóm" các đối tượng "giống" nhau vào "một lớp" dựa trên các đặc trưng dữ liệu của chúng. Hệ thống đánh giá phân lớp các bình luận rút trích được thành 3 nhóm: "tích cực", "tiêu cực" và "trung lập"

Nhóm xây dựng bộ phân lớp dựa trên hai giải thuật là Naïve Bayes và SVM.

3.3.1. Naïve Bayes

Đây là kỹ thuật phân lớp giám sát được đề xuất bởi Thomas Bayes[6]. Phương pháp của Naïve Bayes được sử dụng khá phổ biến trong các lãnh vực tìm kiếm, lọc mail, phân lớp, ... Kỹ thuật này sử dụng xác suất có điều kiện giữa từ và chủ đề để xác định chủ đề của văn bản. Các xác suất này dựa trên việc thống kê sự xuất hiện của từ và chủ đề trong tập huấn luyện. Tập huấn luyện lớp có thể mang lại kết quả khả quan cho Naïve Bayes. Ưu điểm của phương pháp này là đơn giản, tốc độ nhanh, cài đặt không quá phức tạp phù hợp với thời gian cho phép.

Thuật toán gồm 2 giai đoạn huấn luyện và phân lớp:

1. <u>Huấn luyện:</u>

   Tính $P(C_i)$ và $P(x_k|C_i)$
   Công thức tính $P(C_i)$ đã làm trơn Laplace

$$P(C_i) = \frac{|docs_i| + 1}{|total\ docs| + m}$$

   Đầu vào:

   - Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
   - Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.

   Đầu ra:

   - Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

2. <u>Phân lớp:</u>

   Đầu vào:

   - Vector đặc trưng của văn bản cần phân lớp.
   - Các giá trị xác suất $P(C_i)$ và $P(x_k|C_i)$.

   Đầu ra:

   - Nhãn/lớp của văn bản cần phân loại.

3.4. Support Vector Machines (SVM)

Support Vector Machines là một phương pháp máy học do Vladimir Vapnik và các cộng sự [6] xây dựng nên từ những năm 70 của thế kỉ 20. SVMs là bộ phân lớp nhị phân, để áp dụng trong bài toán phân loại đa lớp, một số chiến thuật phân lớp đã được đề xuất, như One-Against-One (OAO), One-Against-Rest (OAR), dựa trên cấu trúc đồ thị (DDAG, ADAG), Half-Against-Half (HAH) và phương pháp phân loại nhiều lớp mờ. Thuật toán phân lớp SVM nhóm chọn là OAO do có nhiều thực nghiệm cho kết quả tương đối tối ưu, được triển khai trên thư viện Mlib của Apache Spark[7].

---

[6] https://bayesian.org/bayes

[7] Apache Spark™ - Lightning-Fast Cluster Computing – http://spark.apache.org

Phương pháp One-Against-One (OAO)

1. Huấn luyện:

   Đầu vào:

   - Các vector đặc trưng của văn bản trong tập huấn luyện (Ma trận MxN, với M là số vector đặc trưng trong tập huấn luyện, N là số đặc trưng của vector).
   - Tập nhãn/lớp cho từng vector đặc trưng của tập huấn luyện.
   - Các tham số cho mô hình SVM: C, γ (tham số của hàm kernel, nhóm sử dụng hàm Gauss)

   Đầu ra:

   - Mô hình SVM (Các Support Vector, nhân tử Lagrange a, tham số b).

2. Phân lớp:

   Đầu vào:

   - Vector đặc trưng của văn bản cần phân lớp.
   - Mô hình SVM

   Đầu ra:

   - Nhãn/lớp của văn bản cần phân loại

4 Kết quả thực nghiệm và đánh giá

Đề tài thực hiện thực nghiệm với thuật toán Native Bayes và SVM kỹ thuật phân đa lớp chiến lược OAO.

Thu thập dữ liệu: Dữ liệu mà đề tài chuẩn bị được thu thập bằng crawler từ diễn đàn công nghệ tinhte.vn[8], trang thương mại điện tử thegioididong.vn[9]. Dữ liệu thu thập về là tập các bình luận đánh giá của người dùng về hai nhóm sản phẩm là *Điện thoại iPhone 6s* (4016 bình luận) và *Điện thoại Xiaomi Redmi Note 2* (566 bình luận), các chủ đề này được crawler lọc theo từ khóa của bài biết đầu tiên (trên tinhte.vn) và theo tên sản phẩm (thegioididong.vn).

Xử lý dữ liệu: Các bài viết, bình luận sau khi thu thập được tiền xử lý và chuẩn hóa. Lọc bỏ các liên kết, lọc bỏ trích dẫn (quote) trong bình luận, gán nhãn dữ liệu. Mỗi bình luận được gán nhãn bằng tay, gồm nhãn: tích *cực (1), tiêu cực (-1), trung lập (0), không liên quan (-2).*

8 http://tinhte.vn
9 http://thegioididong.vn

Toàn bộ hệ xử lý dữ liệu, huấn luyện kiểm tra được xây dựng trên nền tảng của PredictionIO[10] chuyên dùng trong việc xử lý phân tán và xử lý dữ liệu lớn (Big Data).

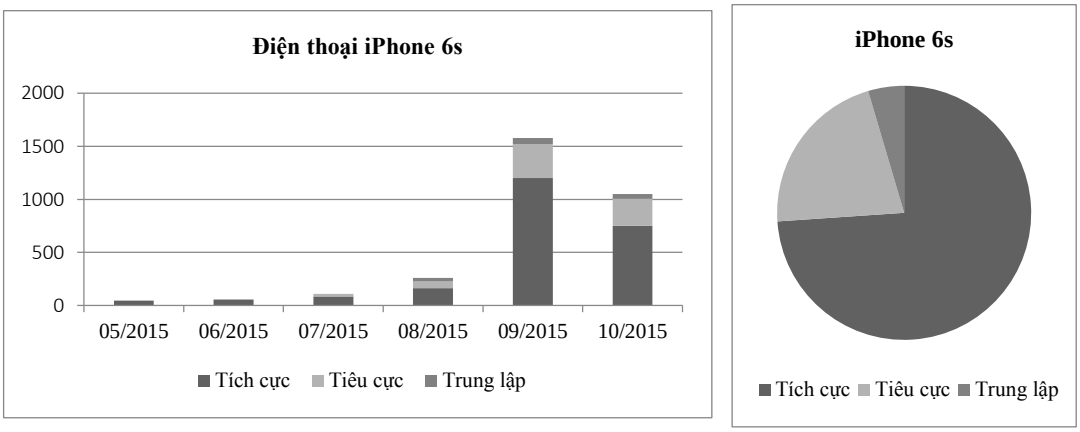
Từ bộ dữ liệu được gán nhãn, chia ra 80% để huấn luyện và 20% để kiểm thử. Nhóm đề tài lần lượt sử dụng thuật toán Naïve Bayes và SVM chiến lược OAO.

Đánh giá hiệu quả phân lớp với độ đo độ chính xác, độ chính xác của thuật toán được xác định bằng cách so sánh kết quả nhãn của bộ dữ liệu kiểm thử sau khi qua bộ phân lớp, với kết quả gán nhãn bằng tay. Kết quả cho thuật toán lần lượt được ghi trong bảng 1.

Bảng 1: Bảng đánh giá phân lớp cho thuật toán Naïve Bayes và SVM (đơn vị: bình luận)

STT	Tên sản phẩm	Huấn luyện	Kiểm thử	Độ chính xác	
				Naïve Bayes	SVM
1	iPhone 6s	3212	803	86.54%	91.15%
2	Xiaomi Redmi Note 2	453	113	82.2%	89.95%

Ta thấy được kết quả phân lớp từ thuật toán SVM cho kết quả tốt hơn nhiều so với thuật toán Naïve Bayes. Từ kết quả này ta có được biểu đồ báo cáo trực quan như hình 1.



Hình 2: Biểu đồ thống kê mức độ quan tâm của sản phẩm điện thoại iPhone 6s theo từng mốc thời gian

[10] PredictionIO (https://prediction.io) - An open-source machine learning server for developers

5 Kết luận

Trong bài viết này nhóm đề tài đã trình bày phương pháp rút trích và phân loại bình luận về sản phẩm điện tử trên Internet, từ kết quả phân loại đó thống kê báo cáo được ý kiến của người dùng về sản phẩm ấy như thế nào, biểu diễn trực quan độ quan tâm theo thời gian. Phương pháp phân loại được triển khai bằng phương pháp máy học, dựa trên hai thuật toán Native Bayes và SVM (chiến lược đa lớp OAO), từ kết quả đánh giá cho thấy thuật toán SVM cho hiệu quả phân lớp cao hơn.

6 Tài liệu tham khảo

1. Trần Cao Đệ, Phạm Nguyên Khang "Phân loại văn bản với máy học vector hỗ trợ", Tạp chí Khoa học 2012:21a 52-63
2. A. Mudinas, D. Zhang, and M. Levene. (2012), "Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis"
3. B. Pang and L. Lee. (2008), "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
4. B. Pang, L. Lee, and S. Vaithyanathan. (2002), "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86.
5. Binh Thanh Kieu and Son Bao Pham. (2010), "Sentiment Analysis for Vietnamese".
6. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
7. Haseena Rahmath P and Tanvir Ahmad. (2014), "Sentiment Analysis Techniques - A Comparative Study", IJCEM International Journal of Computational Engineering & Management, Vol. 17 Issue 4, July 2014.
8. Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu and Bing Liu. (2011), "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis"
9. Lê Hồng Phương, Nguyễn Thị Minh, Azim Roussanaly Huyền, and Hồ Tường Vinh. "A Hybrid Approach to Word Segmentation of Vietnamese Texts." Proceedings of the 2nd International Conference on Language and Automata Theory and Applications. 2008.
10. M. Hu and B. Liu. (2004), "Mining and Summarizing Customer Reviews".
11. M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez and L.A. Ureña-López. (2011), "Experiments with SVM to classify opinions in different domains", Expert Systems with Applications 38 (2011) 14799–14804.
12. Peter D. Turney. (2002), "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised 4 Classification of Reviews", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.
13. R. Prabowo and M. Thelwall. (2009), "Sentiment analysis: A combined approach", Journal of Informetrics, vol. 3, pp.143-157, 2009.