TEXT MINING TRONG PHÂN TÍCH PHẢN HỒI KHÁCH HÀNG

N.Q.Chính
Khoa Toán Kinh Tế
Đại học Kinh tế Quốc dân
1200655@st.neu.edu.vn

ABSTRACT

Đi kèm với tốc độ phát triển nhanh chóng của các nền tảng thương mại điện tử và số lượng lớn phản hồi của khách hàng, là nhu cầu cấp thiết về các phương pháp phân tích dữ liệu văn bản và một bộ khung để quản lý, lưu trữ và xử lý dữ liệu lớn một cách hiệu quả.

1 Giới thiệu

Hiện nay với sự bùng nổ của thời đại công nghệ số, đi kèm với đó là lượng lớn dữ liệu được tạo ra với tốc độ chóng mặt. Theo như ước tính của IBM vào năm 2012, mỗi ngày 2.5 exabytes (2.5 tỷ gigabytes) dữ liệu được tạo ra và 90% dữ liệu hiện nay trên thế giới được tạo ra vỏn vẹn trong giai đoạn 2010-2012 [1]. Trong đó theo một thống kê trong bài báo của Cloudera với tựa đề "The Rise of Unstructed Data"[2], trong khi lượng dữ liệu có cấu trúc gia tăng với tốc độ 12% một năm, thì dữ liệu phi cấu trúc đang tăng một lượng từ 55% đến 65% hằng năm. Qua đó cho chúng ta thấy lượng dữ liệu mà chúng ta tạo ra lớn và tăng trưởng nhanh đến dường nào, đặc biệt đối với dữ liệu phi cấu trúc. Vì thế việc phân tích và lưu trữ để tận dụng một lượng dữ liệu lớn với tốc độ nhanh là điều cấp thiết đối với sự phát triển kinh tế và xã hội. Để đáp ứng nhu cầu cấp thiết ấy chúng tôi đề xuất giải pháp Text Mining để khai thác và thống kê cảm xúc của khách hàng từ dữ liệu phản hồi về sản phẩm. Giải pháp này nhắm đến mục tiêu khai thác và trích xuất thông tin về sản phẩm từ phản hồi của khách hàng một cách tự động nhằm cung cấp cho nhà quản lý sản phẩm thông tin và dữ liệu cần thiết phục vụ cho việc hỗ trợ đưa ra quyết định. Cấu trúc của bài nghiên cứu sẽ được chia như sau: Phần 3 và 4 sẽ giới thiệu về cơ sở lý thuyết và cách huấn luyện của các mô hình chủ đề và mô hình phân tích cảm xúc. Sau đó cơ chế hoạt động của hệ thống sẽ được bàn luận tại phần 5, cuối cùng chúng tôi trình bày kết quả thử nghiệm và đưa ra kết luận lần lượt tại phần 6 và 7.

2 Các nghiên cứu liên quan

3 Mô hình chủ đề

Mô hình hóa chủ đề (Topic Modelling) là một phương pháp phân tích văn bản được sử dụng nhiều trong các nghiên cứu về khoa học xã hội, con người và nhiều hơn thế nữa [3]. Các mô hình chủ đề cung cấp một quá trình phân tích tự động chuyển một hoặc nhiều tệp văn bản trở thành một tập hợp các chủ đề nằm trong tệp văn bản. Lớp mô hình này bao gồm các thuật toán phi giám sát (Unsupervised-learning) như VLDA [4], LDA [5], GSM [6] không cần có dữ liệu được dán nhãn. Trong bài nghiên cứu này chúng tôi sẽ sử dụng thuật tuán LDA làm mô hình phân tích chủ đề. Mô hình chủ đề có hai vai trò trong hệ thống của chúng tôi, thứ nhất trích xuất ra các đặc tính sản phẩm được ẩn trong trong tệp dữ liệu huấn luyện, thứ hai nhận câu nhận xét của khách hàng sau đó phân loại câu nhận xét với đặc tính sản phẩm tương ứng.

Sau đây là phần giải thích ý nghĩa một số kí hiệu và thuật ngữ mà chúng tôi sẽ sử dụng tại phần này. Lưu ý khi chúng tôi nhắc đến "dữ liệu"hay "văn bản"sẽ được hiểu là các câu nhận xét của khách hàng về sản phẩm. "Tệp văn bản"hay "tệp dữ liệu"là một tập hợp nhiều câu nhận xét. "Chủ đề của câu nhận xét"hay "chủ đề của dữ liệu"hay "đặc tính của sản phẩm"sẽ có ý nghĩa tương đương, ví dụ câu nhận xét sau: "Ảnh của máy chụp rất đẹp", chủ đề của câu nhận xét hay đặc tính của sản phẩm được nêu trong câu là chất lượng của ảnh của máy. Sau đây là một số lưu ý về kí hiệu:

- Một từ sẽ được coi là đơn vị cơ bản của dữ liệu rời rạc, các từ sẽ nằm trong một thư viện từ diển có kích cỡ V 1,...,V. Chúng tôi sử dụng các vector cơ sở có một thành phần bằng một, tất cả còn lại bằng không làm đại diện

cho từng từ trong thư viện từ V, vị trí của thành phần bằng 1 sẽ tương ứng với vị trí của từ đó trong từ điển V. Giả sử từ thứ n nằm trong từ điển V sẽ được đại diện bởi một vector $\vec{w}_n \in \mathbb{R}^V$ trong đó $w^n = 1$ và $w^u = 0$ với mọi $u \neq v$.

- Một câu nhận xét hay một văn bản là chuỗi N các từ kí hiệu bởi $W = (w_1, w_2, ..., w_N)$ trong đó w_n là từ thứ n ở trong văn bản.
- Một tệp tệp nhận xét hay một tệp văn bản là tập hợp M câu nhận xét hoặc văn bản được kí hiệu bởi $D = (W_1, W_2, ..., W_M)$

3.1 Mô hình phân bố Dirichlet tiềm ẩn

Mô hình phân bố Dirichlet tiềm ẩn (Latent Dirichlet Allocation - LDA) là một mô hình học máy phi giám sát (Unsupervised Learning) hoạt động dưới cơ chế của một mô hình xác suất tạo sinh (Generative Model) dành cho dữ liệu rời rạc như dữ liệu văn bản, được sử dụng để khai thác các chủ đề tiềm ẩn trong một tệp văn bản không được gán nhãn. Ý tưởng của mô hình LDA là mỗi văn bản được đại diện bởi sự kết hợp ngẫu nhiên của các chủ đề, trong đó mỗi chủ đề được mô hình hóa bởi phân phối xác suất của các từ. Mục tiêu của LDA là mô hình hóa phân bố xác suất đồng thời của văn bản và chủ đề $p(W, z)$, khi cần phân loại văn bản thuộc chủ đề nào chúng ta sử dụng công thức Bayes. LDA đưa ra giả thuyết về các một văn bản W nằm trong tệp D như sau:

1. Chọn $N \sim Poisson(\xi)$.
2. Chọn $\theta \sim Drichlet(\alpha)$.
3. Với mỗi từ w_n trong N từ:
   (a) Chọn một chủ đề $z_n \sim Multinomial(\theta)$.
   (b) Chọn một chữ w_n từ $p(w_n|z_n, \beta)$, là một phân phối đa thức (Multinomial Distribution) có điều kiện là chủ đề z_n.

Quá trình trên chính là quá trình một văn bản bất kì được sinh ra dụa trên giả thiết của mô hình LDA. N ở đây là số lượng từ trong một văn bản, trong bài báo gốc của LDA tác giả có đưa ra gợi ý rằng N không nhất thiết phải được lấy từ phân phối Poisson, mà hoàn toàn có thể thay thế bằng các phân phối khác tùy thuộc vào dữ liệu của bài toán. Sau khi có tham số N là đọ dài của văn bản, chúng ta chọn tham số θ là một biến ngẫu nhiên có số chiều bằng số lượng chủ đề kí hiệu là k, và θ thỏa mãn $\theta_i \geq 0$, $\Sigma_{i=1}^k \theta_i = 1$. θ sẽ sử dụng làm tham số cho phân phối đa thức để chọn chủ đề cho văn bản. Trong đó tham số θ tuân theo phân phối Dirichlet:

$$p(\theta|\alpha) = \frac{\Gamma(\Sigma_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}, \tag{1}$$

$$\Gamma(x) = (x-1)!\ ,\ x > 0, \tag{2}$$

trong đó tham số α là vector có chiều bằng k, mỗi thành phần $\alpha_i > 0$, $\Gamma(x)$ là hàm Gamma có công thức số (2). Tại sao tại đây tác giả của LDA lại sử dụng phân phối Dirichlet để mô hình hóa θ. Đầu tiên do sẽ được sử dụng làm tham số cho phân phối đa thức ở giai đoạn sau nên θ phải thỏa mãn điều kiện $\theta_i \geq 0$, $\Sigma_{i=1}^k \theta_i = 1$ và phân phối Dirichlet thỏa mãn được những điều kiện trên của θ. Hơn nữa phân phối Dirichlet thuộc lớp phân phối mũ, là một thống kê đủ cho tập số hữu hạn, các tính chất giúp quá trình ước lượng tham số và tính toán về sau dễ dàng hơn.

Sau khi chọn N và θ, quá trình tạo ra văn bản đi qua các bước tiếp như sau, đầu tiên chúng ta sẽ khởi tạo từng từ một do văn bản là chuỗi N các từ nối tiếp với nhau. Để khởi tạo một từ, đầu tiên ta khởi tạo một chủ đề z_n cho từ đó sử dụng phân phối nhị thức với tham số θ đã được khởi tạo ở bước trên. Tiếp theo khởi tạo giá trị cho từ w_n bằng phân phối đa thức có điều kiện, sử dụng chủ đề z_n đã được khởi tạo để làm điều kiện cho w_n. Lưu ý α và β là hai siêu tham số của mô hình LDA không ước lượng hay mô hình hóa bằng phân phối như N và θ, tác động của hai siêu tham số này đến độ chính xác của mô hinh sẽ được đưa ra tại phần 6 của bài nghiên cứu, tại đó chúng tôi sẽ đưa ra một số khuyến nghị về việc tinh chỉnh sao cho đạt được kết quả dự báo tốt nhất.

4 Mô hình phân tích cảm xúc

5 Cơ chế hoạt động

Hệ thống Text Mining của chúng tôi gồm hai thành phần chính: mô hình chú đề và mô hình phân tích cảm xúc. Mô hình chủ đề LDA (Latent Dirichlet Allocation) [5] kết hợp với kĩ thuật Gibbs Sampling sẽ được sử dụng để trích xuất chủ đề

dữ liệu, chúng tôi sẽ sử dụng mô hình RWKV [7] được huấn luyện sử dụng phương pháp chắt lọc tri thức [8] làm mô hình phân tích cảm xúc. Khi đưa vào hoạt động hệ thống sẽ đưa dữ liệu qua các bước như sau:

- Bước một: Dữ liệu được đưa qua các bước tiền xử lý bao gồm các bước làm sạch và chuyển về định dạng phù hợp. Do dữ liệu của bài toán ở dưới dạng văn bản các bước làm sạch dữ liệu sẽ bao gồm: chuẩn hóa văn bản, loại bỏ dấu câu và loại bỏ chữ số. Sau khi đi qua các kĩ thuật làm sạch dữ liệu, dữ liệu văn bản được đưa về dạng số trở để định dạng phù hợp cho các bước kế tiếp.
- Bước hai: Mô hình chủ đề lấy đầu ra của bước một để phân tích và đưa ra phân phối xác suất về chủ đề tương ứng của câu. Sau đó dựa vào phân phối xác suất dữ liệu sẽ được gán nhãn với một hoặc nhiều chủ đề tương ứng.
- Bước ba: Đầu ra của bước một sẽ được đưa vào mô hình phân tích cảm xúc để phân loại phản hồi của khách hàng. Sẽ có ba mức độ phân loại: Tích cực, trung hòa, tiêu cực.
- Bước bốn: Cuối cùng đầu ra của bước hai và ba cùng với câu phản hồi tương úng sẽ được lưu vào cơ sở dữ liệu, phục vụ cho mục đích truy vấn và phân tích của nhà quản lý sản phẩm.

Nhà quản lý sản phẩm khi muốn tìm hiểu về một sản phẩm nào đó sử dụng hệ thống của chúng tôi sẽ cần cung cấp những dữ liệu sau đây. Đầu tiên là dữ liệu phản hồi, comments của khách hàng về sản phẩm của mình, số lượng mẫu phản hồi sẽ tùy thuộc vào độ tin cậy, chính xác về mặt thông tin mà nhà quản lý mong muốn, dữ liệu càng nhiều và chất lượng càng cao thì kết quả thống kê đưa ra càng tốt. Tiếp theo người phân tích cần cung cấp số lượng đặc tính của sản phẩm mà họ muốn phân tích tùy vào sản phẩm và mong muốn của người phân tích, thông thường con số này nằm trong khoảng từ 1 đến 10. Sau khi mô hình LDA đã được huấn luyện, mô hình sẽ đưa ra thống kê các từ khóa liên quan đến từng đặc tính tương ứng mà người phân tích đã cung cấp, dựa vào dữ liệu thống kê mà người phân tích sẽ đặt tên cho từng đặc tính sau đó mô hình sẽ sử dụng tên đã được đặt cho đặc tính để gán nhãn cho các câu về sau.

6 Kết quả thử nghiệm

7 Kết luận

Tài liệu

[1] National Institute of Technology Silchar. Big data analytics - what is that?

[2] Daniel Valdez Balderas. The rise of unstructured data, 2021. November 15, 2021.

[3] John W. Mohr and Petko Bogdanov. Introduction—topic models: What they are and why they matter. *Poetics*, 41(6):545–569, 2013. Topic Models and the Cultural Sciences.

[4] Zhongyuan Tian, Harumichi Yokoyama, and Takuya Araki. Parallel latent dirichlet allocation using vector processors. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1548–1555, 2019.

[5] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research 3 (2003) 993-1022*, 2003.

[6] Phil Blunsom Yishu Miao, Edward Grefenstette. Discovering discrete latent topics with neural variational inference. *arXiv:1706.00359*, 2018.

[7] Quentin Anthony Alon Albalak Samuel Arcadinho Huanqi Cao Xin Cheng Michael Chung Matteo Grella Kranthi Kiran GV Xuzheng He Haowen Hou Przemyslaw Kazienko Jan Kocon Jiaming Kong Bartlomiej Koptyra Hayden Lau Krishna Sri Ipsit Mantri Ferdinand Mom Atsushi Saito Xiangru Tang Bolun Wang Johan S. Wind Stansilaw Wozniak Ruichong Zhang Zhenyuan Zhang Qihang Zhao Peng Zhou Jian Zhu Rui-Jie Zhu Bo Peng, Eric Alcaide. Rwkv: Reinventing rnns for the transformer era. *arXiv:2305.13048*, 2023.

[8] Linqing Liu Lili Mou Olga Vechtomova Jimmy Lin Raphael Tang, Yao Lu. Distilling task-specific knowledge from bert into simple neural networks. *arXiv:1903.12136*, 2019.