Aspect-Based Sentiment Analysis in Vietnamese Restaurant

Viet-Anh Tran[1,3,4,5,6], Thi-Thu-Hang Le[2,3,4,5,6], Hoang Bao-Long Nguyen[2,3,4,5,6], and Nguyen-Phuong Hoang[1,3,4,5,6]

[1] IE212.M11.CNCL
[2] IE212.M12.CNCL
[3] Faculty of Information Science and Engineering
[4] University of Information Technology, Ho Chi Minh City, Vietnam
[5] Vietnam National University, Ho Chi Minh City, Vietnam
[6] {18520253,18520274,18521039,18521270}@gm.uit.edu.vn

Tóm tắt nội dung Trong bài báo khoa học này nhóm xây dựng một ứng dụng liên quan đến việc đánh giá comment của khách hàng về phía cửa hàng, để làm được điều đó nhóm sử dụng bộ dữ liệu data do nhóm tự xây dựng với độ đồng thuận là 85%. Bộ dữ liệu này được thu thập trên các trang mạng uy tín chuyên về review ẩm thực nhằm có thể tạo ra được một bộ dữ liệu khách quan, đúng chuẩn nhất có thể. Sau khi có được bộ dữ liệu, nhóm sử dụng PySpark là chủ yếu với các thư viện hỗ trợ chính như BigDL và Analytics zoo. Kết quả đạt được là đã tạo ra được một ứng dụng vớ nhiều model khác nhau để có được kết quả tổng quan nhất nhưng độ chính xác cao nhất là 0.71 với model RNN và CountVectorizer.

Keywords: ABSA · BigDL · Analytics-Zoo

1 Giới thiệu

Ngày nay, với sự phát triển của công nghệ, các trang mạng xã hội như Facebook, Tiktok, Twitter,... đã được hình thành, cung cấp cho mọi người những nơi để giải trí, kết bạn hay chia sẻ thông tin với nhau. Thông tin thường xuyên được chia sẻ trên mạng xã hội thường là những lời giới thiệu, bình luận hoặc chia sẻ kinh nghiệm. Từ đó, việc quảng bá thương hiệu là một lợi thế nếu người bán biết tận dụng mạng xã hội, cho phép người dùng tìm hiểu thêm về cửa hàng của họ. Ngoài ra, trên mạng xã hội, các thực khách cũng thường viết những nhận xét hoặc đánh giá về những địa điểm họ đã ăn để chia sẻ với mọi người. Đối với khách hàng, những bài viết này giúp thực khách hiểu được họ sẽ dùng bữa ở đâu để cân nhắc có nên dành thời gian và tiền bạc cho trải nghiệm hay không. Còn đối với người bán, họ có thể cải thiện cửa hàng của mình bằng những đánh giá này.

Từ vấn dề đó mà chúng tôi đã quyết định chọn đề tài "Xác định ý kiến (cảm nhận) của khách hàng ở các khía cạnh thông qua các đánh giá về nhà hàng"nhằm giúp cho người dùng có thể có một cái nhìn tổng quan nhất về nhà hàng. Đồng thời, có thể giúp nhà hàng cải thiện chất lượng của chính mình

2 Công trình nghiên cứu liên quan

Bài báo Analysis and Performance Evaluation of Deep Learning on Big Data của Matteussi và các cộng sự đã nói về việc kết hợp mô hình DeepLearning với khả năng xử lý của BigDL nhằm mục đích tiến hành phân tích và đánh giá hiệu suất các ứng dụng DeepLearning trong BigDL.Kết quả cho thấy tính khả thi của xử lý phân tán, với tốc độ tăng lên đến 8 lần và độ chính xác bị giảm đi 5% trong trường hợp tốt nhất. Bài báo LSTM based time của Guoqiong Song và các cộng sự đã nói về cách sử dụng Analytics Zoo cho BigDL. Việc sử dụng LSTM với các layer của BigDL giúp các dữ liệu phân tính có thể hoạt động trên mô hình DeepLearning của Tensorflow. Kết quả cho thấy tiết kiệm chi phí cho việc đào tạo và bảo trì dự đoán các dữ liệu phân tán trên spark .

3 Bộ dữ liệu

Chúng tôi tạo một tập dữ liệu phân loại cảm xúc theo khía cạnh (ABSA) mới bằng tiếng Việt. Quy trình tạo ra tập dữ liệu của chúng tôi được mô tả như sau. Đầu tiên, chúng tôi thu thập ý kiến từ các trang web chuyên về review, đánh giá ẩm thực tại Việt Nam (xem phần 3.1). Tiếp theo, chúng tôi xây dựng một hướng dẫn cho những người gán nhãn để xác định các khía cạnh và ý kiến trong các bình luận, để giúp họ hiểu hơn và gán nhãn dữ liệu chính xác hơn (xem phần 3.2). Sau đó, tạo và chỉnh sửa hướng dẫn gán nhãn để đảm bảo độ đồng thuận của tập dữ liệu trên 80% trước khi thực hiện một cách độc lập (xem phần 3.3). Cuối cùng, chúng tôi phân tích về tập dữ liệu để giúp hiểu hơn về tập dữ liệu này (xem phần 3.4).

3.1 Thu thập dữ liệu

Chúng tôi đã thu thập các nhận xét, đánh giá của khách hàng trên các trang web chuyên về review về ẩm thực như Foody, Loship, Tripadvisor và các hội nhóm review trên Facebook. Để đám bảo các thông tin thu được đa dạng và phong phú, chúng tôi đã giữ lại toàn bộ các bình luận mà chúng tôi thu thập được.

3.2 Những khía cạnh trong tập dữ liệu

Trong các bài đánh giá mà chúng tôi thu thập được, người dùng nhận xét trên nhiều khía cạnh rõ ràng hoặc ẩn ý về nhiều khía cạnh như giá cả, giá cả, chất lượng, dịch vụ và vân vân. Tập dữ liệu bao gồm 10.117 phản hồi với bảy khía cạnh: general, price, quality, service, stylefood, location, background và mỗi khía cạnh có một trong ba ý kiến: positive (nhận xét tích cực, hài lòng), negative (nhận xét tiêu cực, phàn nàn) và neutral (nhận xét trung lập, hoặc có cả tích cực và tiêu cực). Bảng 1 tóm tắt bảy khía cạnh trong hướng dẫn. Đối với những nhận xét không liên quan đến khía cạnh nào hoặc không đánh giá sản phẩm, chúng tôi không gán nhãn cho những trường hợp đó.

Bảng 1. Các khía cạnh trong tập dữ liệu

Nhãn	Định nghĩa
General	Xét về tổng quan nhà hàng hoặc đề cập tới một loại thực thể chung, không rõ ràng, thể hiện tính nói chung về nhà hàng. Hoặc là lời hứa quay lại hay không, tần suất ghé thăm, đặc trưng quán, độ nổi tiếng, cảm nhận về nhà hàng.
Price	Đề cập tới giá thức ăn, đồ uống, phí dịch vụ được cung cấp bởi nhà hàng, hoặc giá nhà hàng nói chung. Chỉ nói về giá thì nhãn 0.
Quality	Chất lượng đồ ăn, đồ uống
Service	Đề cập tới thái độ phục vụ, cách phục vụ khách hàng, các chương trình, dịch vụ cho khách hàng
FoodStyle	Cách bày trí thức ăn, đồ ăn kèm, đồ uống hoặc nói về tính đa dạng về đồ ăn thức uống phục vụ trong nhà hàng.
Location	Đề cập tới vị trí của nhà hàng
Background	Không gian của nhà hàng, có đông khách hay không

3.3 Quy trình gán nhãn

Ba giai đoạn của gán nhãn được tiến hành như sau. Để bắt đầu, chúng tôi đào tạo người gán nhãn với các nguyên tắc và lấy ngẫu nhiên khoảng 200-250 bình luận đánh giá trong tập dữ liệu để ghi hướng dẫn gán nhãn, sau đó tính toán độ đồng thuận cho những dữ liệu được gán nhãn đó. Đối với các trường hợp không đồng ý, người gãn nhãn quyết định nhãn cuối cùng bằng cách thảo luận và có một cuộc thăm dò ý kiến. Người gán nhãn dành năm vòng đào tạo để đạt được độ đồng thuận cao trên 80% trước khi thực hiện gán nhãn dữ liệu một cách độc lập. Độ đồng thuận giữa những người gán nhãn được tính toán dựa theo hệ số Cohen's Kappa [3], độ đồng thuật đạt 84.2%.

Cuối cùng, tập dữ liệu của chúng tôi được chia ngẫu nhiên thành hai tập: train và test theo tỉ lệ 8:2.

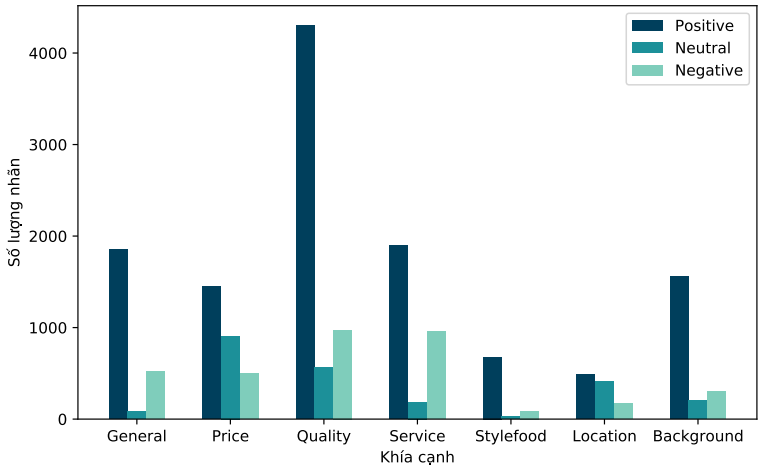
3.4 Phân tích dữ liệu

Hình 1 trình bày sự phân bố của bảy danh mục khía cạnh trong tập dữ liệu của chúng tôi. Mọi người có xu hướng đánh giá chất lượng (QUALITY) của nhà hàng. Khách hàng thường xuyên chú ý đến các khía cạnh liên quan đến nhu cầu của họ như PRICE, SERVICE và BACKGROUND. Thống kê tập dữ liệu của chúng tôi được trình bày trong Bảng 2.

Tập dữ liệu của chúng tôi gồm 18.098 nhãn trên 10.117 bình luận. Thông qua phân tích của chúng tôi, tập dữ liệu phân bố không đồng đều giữa các nhãn. Nhãn tích cực (positive) chiếm số lượng nhiều nhất trong các nhãn. Tiếp theo là tiêu cực (negative) và trung tính (neutral).

Bảng 2. Thống kê tổng quan về các nhãn trong tập dữ liệu

	General	Price	Quality	Service	StyleFood	Location	Background	Tổng nhãn
Positive	1856	1449	4302	1895	678	486	1559	12225
Neutral	82	909	567	187	32	413	200	2390
Negative	523	495	973	958	89	172	302	3512

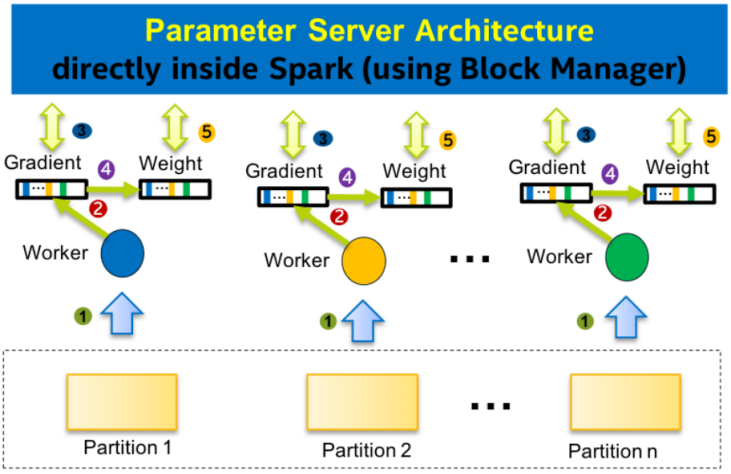


Hình 1. Sự phân bố của 7 danh mục khía cạnh trong tập dữ liệu

4 Phương pháp

Sau khi có một bộ dữ liệu theo chuẩn dữ liệu, chúng tôi tiến hành tìm hiểu phương pháp tiếp cận để có thể đánh giá bình luận qua các phía cạnh. Ở các hệ thống tệp phân tán Apache Hadoop (HDFS), Apache Storm / Kafka và các thành phần khác), trong đó có thể phức tạp và dễ xảy ra lỗi tiến trình. Nhằm có thể xây dựng một mô hình học sâu đáp ứng cần phải đầu tư vào việc xây dựng các lớp riêng biệt để tận dụng các khả năng của mạng neural trên nền tảng spark.

Tại Intel đã xây dựng các ứng dụng học sâu và AI trên Apache Spark. Để hợp lý hóa việc triển khai và phát triển end-to-end, Intel đã phát triển Analytics Zoo , một nền tảng phân tích + AI hợp nhất, kết hợp liền mạch các chương trình spark, và sử dụng BigDL thành một Pipeline tích hợp, có thể mở rộng một cách rõ ràng cho Apache Hadoop / Spark lớn để đào tạo hoặc dự đoán trên dữ liệu phân tán.



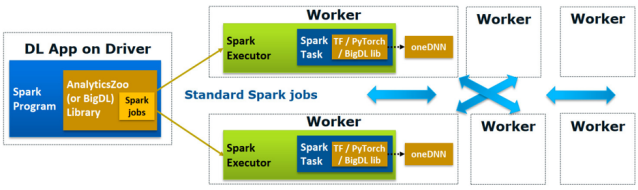
Hình 2. Quy trình hoạt động của BigDL

4.1 Giới thiệu BigDL và Analytics-Zoo

BigDL [5] là một thư viện học tập sâu phân tán được tạo và mở nguồn bởi Intel. Thư viện được thiết kế từ đầu để chạy tự nhiên trên Apache spark và do đó cho phép các kỹ sư dữ liệu và nhà khoa học viết các ứng dụng học sâu như các chương trình spark tiêu chuẩn mà không phải quản lý rõ ràng các tính toán phân tán.

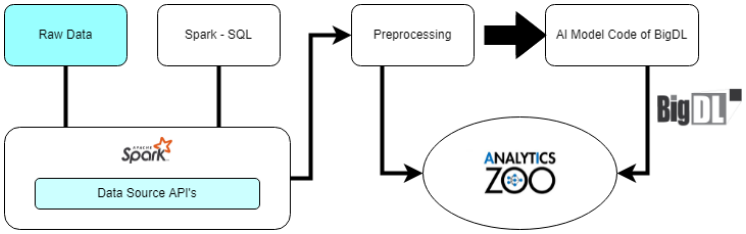
BigDL giúp dữ liệu cần thiết đào tạo như một SparkJob. và mỗi tác vụ của spark chạy trên cùng một mô hình trên một tập dữ liệu như hình 2.

Analytics-Zoo [1] là nền tảng dữ liệu lớn mã nguồn mở và có nhiều tính năng để mở rộng thành dữ liệu lớn. Xây dựng các ứng dụng học sâu end-to-end cho dữ liệu lớn phân tán TensorFlow trên spark. Hình 3 có thể cho thấy việc chạy dữ liệu và mô hình phân tán trên nền tảng Spark và có thể chạy song song đồng bộ các tác vụ spark giúp dữ liệu phân tán có kết quả tốt hơn.



Hình 3. Quy trình hoạt động của Analytics-Zoo

4.2 Quy trình hoạt động của BigDL và Analytics-Zoo



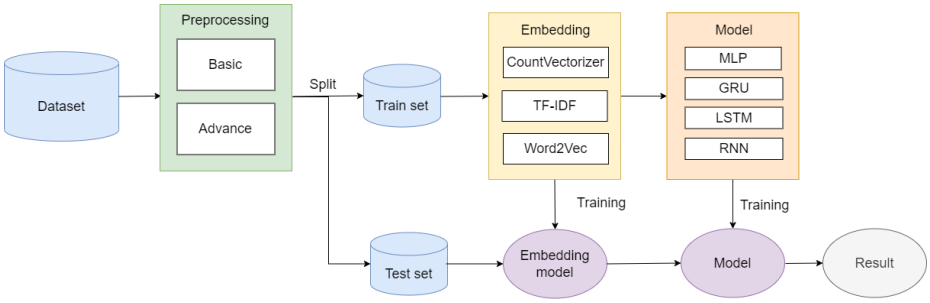
Hình 4. Quy trình của BigDL và Analytics-Zoo

Hình 4 là quy trình hoạt động sử dụng dữ liệu phân tán từ Apache spark và kết hợp với mô hình từ thư viện BigDL. Analytics Zoo cung cấp NNEstimator cho đào tạo mô hình với spark Dataframe, cung cấp API cấp cao để đào tạo một mô hình BigDL với Apache spark Estimator và Transfomer , do đó người dùng có thể đưa Analytics Zoo vào một Pipeline ML một cách thuận tiện.

NNEstimator hỗ trợ các loại dữ liệu nhãn và tính năng khác nhau thông qua Preprocessing. Trong quá trình fit, NNEstimator sẽ trích xuất dữ liệu và nhãn từ spark DataFrame và chuyển đổi dữ liệu cho mô hình thành BigDL Sample.Kết quả fit sẽ là model NNModel, để dự đoán các dữ liệu spark DataFrame.

5 Thực nghiệm

Tổng quan quá trình thực nghiệm của chúng tôi như hình 5. Việc tiền xử lý dữ liệu và embedding được thực hiện trên spark, việc xây dựng các layer và model được tiến hành trên BigDL [2] và Analytics-Zoo [1]. Chi tiết về từng quá trình chúng tôi sẽ trình bày ở phần sau.



Hình 5. Tổng quan quy trình thực nghiệm

5.1 Tiền xử lý dữ liệu

Với bộ dữ liệu ABSA, định dạng ban đầu của dữ liệu là csv có 8 cột gồm cột bình luận và 7 cột của 7 phía cạnh. Vì các bình luận còn thô sơ nên chúng tôi tạo ra phương pháp tiền xử lý cơ bản và tiền xử lý nâng cao.

- Tiền xử lý cơ bản: Chuẩn hóa các unicode, xử lý các icon và bỏ các kí tự đặc biệt.
- Tiền xử lý sâu: Gồm những phương pháp trong tiền xử lý văn bản và thêm một số phương pháp như: Sửa lỗi chính tả, các từ viết tắt thông dụng. Chuyển các hastag (#bunbohue, #danang,..) thành HASTAG. Chuyển giá trị giá cả (10k, 10 ngàn,...) thành GIÁ. Và cuối cùng sử dụng word tokenizer của thư viện Underthesea để xử lý từ trong câu.

Sau khi xử lý các dữ liệu thừa trong các bình luận, chúng tôi kết hợp bảy cột của bảy phía cạnh thành 1 cột tên label có kiểu dữ liệu là mảng chứa các phần tử float. Các phần tử trong mảng được quy định như sau:

- -2: Không có phía cạnh này trong bình luận đang xét.
- -1: Bình luận được đánh giá tiêu cực ở phía cạnh đang xét.
- 0: Về phía cạnh đang xét, bình luận có thể vừa mang tính tiêu cực và tính cực.
- 1: Bình luận được đánh giá tích cực ở phía cạnh đang xét.

Sau khi kết hợp, có được hai cột gồm: một cột chứa các câu bình luận đánh giá của khách hàng và cột còn lại là mảng chứa các giá trị của khía cạnh.

Chúng tôi tiến hành embedding cột cmt bằng Pyspark thành cột features chứa các vector để chuẩn hóa đầu vào các layer của thư viện BigDL. Các embedding được thực hiện bằng thư viện spark-ml [8], gồm: Count Vectorizer, TF-IDF và Word2Vec [9].

5.2 Huấn luyện mô hình

Chúng tôi xây dựng 4 mô hình học sâu là Recurrent neural network [7] (RNN), Gated recurrent unit [4] (GRU), Long short-term memory [6] (LSTM), và Multilayer perceptron (MLP), với các layer của thư viện BigDL và sử dụng Estimator của Analytics-Zoo để huấn luyện và kiểm tra mô hình. Sau khi có mô hình mong muốn, chúng tôi sử dụng streamlit để có thể dễ dàng sử dụng trên web.

Trong đồ án lần này, chúng tôi đã sử dụng learning_rate là 0.2 và batch_size là 64 cho cả hai quá trình huấn luyện và kiểm tra mô hình. Chúng tôi đã đào tạo mô hình của mình trên 10 epochs và sử dụng hàm MSECriterion() của thư viện Analytics-Zoo làm hàm tiêu chuẩn.

5.3 Kết quả

Sau khi huấn luyện, các mô hình có được kết quả được như bảng 3. Như vậy với tiền xử lý cơ bản mô hình có kết quả cao nhất là MLP với embedding là TF-ID đạt 0.71373. Còn với tiền xử lý sâu, mô hình đạt kết quả cao nhất là RNN với embedding CountVectorizer đạt 0.71912.

Bảng 3. Kết quả đánh giá thực nghiệm

Model	Basic Preprocessing	Advance Preprocessing
CountVectorizer + LSTM	0.65809	0.69018
TF-IDF + LSTM	0.69768	0.69488
Word2Vec + LSTM	0.49345	0.48700
CountVectorizer + GRU	0.67435	0.67430
TF-IDF + GRU	0.67407	0.68654
Word2Vec + GRU	0.53275	0.54193
CountVectorizer + MLP	0.69481	0.70441
TF-IDF + MLP	**0.71373**	0.70637
Word2Vec + MLP	0.54922	0.57423
CountVectorizer + RNN	0.68815	**0.71912**
TF-IDF + RNN	0.68262	0.68003
Word2Vec + RNN	0.56470	0.56491

6 Kết luận

Bài toán được đặt ra ở đầu bài đã có giải pháp khá tốt khi tiền xử lý cơ bản với mô hình MLP with TF-IDF với độ chính xác trung bình là 0.71373, khi tiền xử lý cơ bản với mô hình RNN with CountVectorizer với độ chính xác trung bình là 0.71912. Với độ chính tầm 0.71 thì sẽ còn có thể cải thiện tốt hơn với việc thực nghiệm thay đổi các tham số mô hình và tăng cường thêm dữ liệu để huấn luyện mô hình. Tuy nhiên, vẫn còn nhược điểm thời gian khi embedding dữ liệu và huấn luyện mô hình. Tương lai gần sẽ áp dụng thêm các mô hình khác được hỗ trợ trên spark cũng như BigDL và Analytics-Zoo để giải quyết tốt hơn bài toán Aspect Based Sentiment Analytic.

Tài liệu

1. Analytics zoo documentation. https://analytics-zoo.readthedocs.io/en/latest/, (Accessed on 01/26/2022)
2. Bigdl documentation — bigdl documentation. https://bigdl.readthedocs.io/en/latest/, (Accessed on 01/29/2022)
3. Bhowmick, P.K., Basu, A., Mitra, P.: An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics. pp. 58–65. Coling 2008 Organizing Committee, Manchester, UK (Aug 2008), https://aclanthology.org/W08-1209
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
5. Dai, J.J., Wang, Y., Qiu, X., Ding, D., Zhang, Y., Wang, Y., Jia, X., Zhang, C.L., Wan, Y., Li, Z., Wang, J., Huang, S., Wu, Z., Wang, Y., Yang, Y., She, B., Shi, D., Lu, Q., Huang, K., Song, G.: Bigdl: A distributed deep learning framework for big data. CoRR **abs/1804.05839** (2018), http://arxiv.org/abs/1804.05839

6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (12 1997). https://doi.org/10.1162/neco.1997.9.8.1735
7. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences **79**(8), 2554–2558 (1982). https://doi.org/10.1073/pnas.79.8.2554, https://www.pnas.org/content/79/8/2554
8. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., et al.: Mllib: Machine learning in apache spark. The Journal of Machine Learning Research **17**(1), 1235–1241 (2016)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)