PHÁT HIỆN DANH MỤC KHÍA CẠNH TRÊN MIỀN DỮ LIỆU NHÀ HÀNG SỬ DỤNG SPARK NLP

Nguyễn Đức Trí[1,2,*], Trần Trung Hiếu[1,2,*], Nguyễn Hoàng Nhân[1,2,*], and Phạm Phú Phước[1,2,*]

[1] University of Information Technology
[2] Vietnam National University Ho Chi Minh City
[*] *{18521527, 18520754, 18521176, 18520131}@gm.uit.edu.vn*

Tóm tắt nội dung Bài toán Aspect-Based Sentiment Analysis (ABSA) là một trong những thách thức của lĩnh vực xử lý ngôn ngữ tự nhiên. Trong đó bài toán Aspect Category Detection (ACD) là bài toán con của ABSA. Mục đích bài toán này là xác định các loại khía cạnh dựa trên các bình luận do người dùng phản hồi. Đó có thể được coi là một ứng dụng để khai thác ý kiến người dùng về sản phẩm tiêu dùng. Trên thực tế, lượng dữ liệu từ những câu bình luận tăng nhanh và đây được gọi là dữ liệu lớn đòi hỏi các kỹ thuật xử lý đặc biệt và sức mạnh tính toán cao để thực hiện các tác vụ khai thác dữ liệu cần thiết. Mục tiêu của bài báo này là thực hiện được bài toán ACD bằng nền tảng dữ liệu lớn cụ thể là Apache Spark.

Keywords: Spark · ABSA · Aspect based Sentiment Analysis

1 Giới thiệu

Trong những năm gần đây với sự phát triển công nghệ 4.0 và sự nổi lên những trang web thương mại điện tử, mạng xã hội, các blog thì giờ đây việc mọi người bình luận đánh giá sản phẩm, bày tỏ quan niệm trên các trang này là việc rất dễ dàng và thường xuyên qua Internet. Chính vì thế các doanh nghiệp nếu muốn cạnh tranh tốt trong thị trường thì việc nắm bắt thông tin nhanh về nhu cầu khách hàng qua những câu bình luận, đánh giá là một trong những việc làm cần thiết vì việc đó đóng một vai trò quan trọng trong việc đo lường doanh số bán hàng và cải thiện các chiến lược tiếp thị kinh doanh. Ngoài ra các bình luận sẽ là một nguồn dữ liệu giá trị của doanh nghiệp, tổ chức trong việc khai thác, phân tích dữ liêu khách hàng từ đó biết được xu hướng nhu cầu khách hàng mang lại những giá trị, lợi nhuận quan trọng trong việc phát triển mở rộng thị trường. Bài toán ABSA được chia thành ba nhiệm vụ phụ khác nhau, trong đó có Aspect Category Detection, Opinion Term Expression và Sentiment Polarity Classification. Mục tiêu chính trong bài báo này là bài toán Aspect Category Detection trên dữ liệu tiếng Anh về nhà hàng. Bài toán

Aspect Category Detection là phát hiện một hoặc nhiều khía cạnh trong danh sách danh mục được xác định trước của câu bình luận. Với câu đánh giá về nhà hàng: "The food here is rather good, but only if you like to wait for it". Đầu ra sẽ là ['FOOD#QUALITY', 'SERVICE#GENERAL']. Như vậy ta thấy việc phân tích chi tiết các khía cạnh sẽ giúp cho tổ chức doanh nghiệp biết được chính xác các khía cạnh cần được cải thiện, giúp cho chất lượng sản phẩm dịch vụ được nâng cao.

2 Công trình liên quan

Với sự phát triển của học sâu có nhiều mô hình khác nhau được đề xuất để giải quyết ACD. Zhou và cộng sự [1] đã trình bày một cách tiếp cận cho nhiệm vụ ACD dựa trên biểu diễn từ bán giám sát cùng với việc trích xuất đặc trưng qua mạng nơ-ron thông qua các các vectơ từ. Cuối cùng, thuật toán phân loại Logistic Regression sử dụng các đặc trưng trên để dư đoán khía cạnh. Ngoài ra tác giả Xue và cộng sự [2] trình bày một mạng nơ-ron dựa trên lớp BiLSTM kết hợp với lớp CNN cho nhiệm vụ xác định khía cạnh và từ khía cạnh. Gần đây, BERT là kiến trúc được đánh dấu là sự khởi đầu của cộng đồng ngôn ngữ tự nhiên. Ramezani và cộng sự [3] đã trình bày một nghiên cứu về trích xuất ngữ cảnh dựa trên các mô hình BERT để cải thiện model cho nhiệm vụ ACD

Trong cuộc thi (SemEval) 2014, bài toán ABSA đã được giới thiệu đầu tiên cho tiếng Anh cấp độ câu được đánh giá cho hai miền dữ liệu nhà hàng và khách sạn. Điểm F1-score tốt nhất trên hai nhiệm vụ: khía cạnh phát hiện và phân cực khía cạnh đã đạt được bởi NRC-Canada dựa trên năm SVM nhị phân với các tính năng dưới dạng n-gram, từ vựng học được từ dữ liệu YELP cho mỗi khía cạnh cho miền nhà hàng. Nhiệm vụ Aspect Category Detection và Aspect-Sentiment Classification này những nhiêm vụ chính trong SemEval 2014, các nghiên cứu trước đây đều dựa trên kỹ thuật học có giám sát với thuật toán SVM là thuật toán mạnh mẽ được tiếp cận theo hướng mã hóa nhị phân. [15]

3 Phương pháp

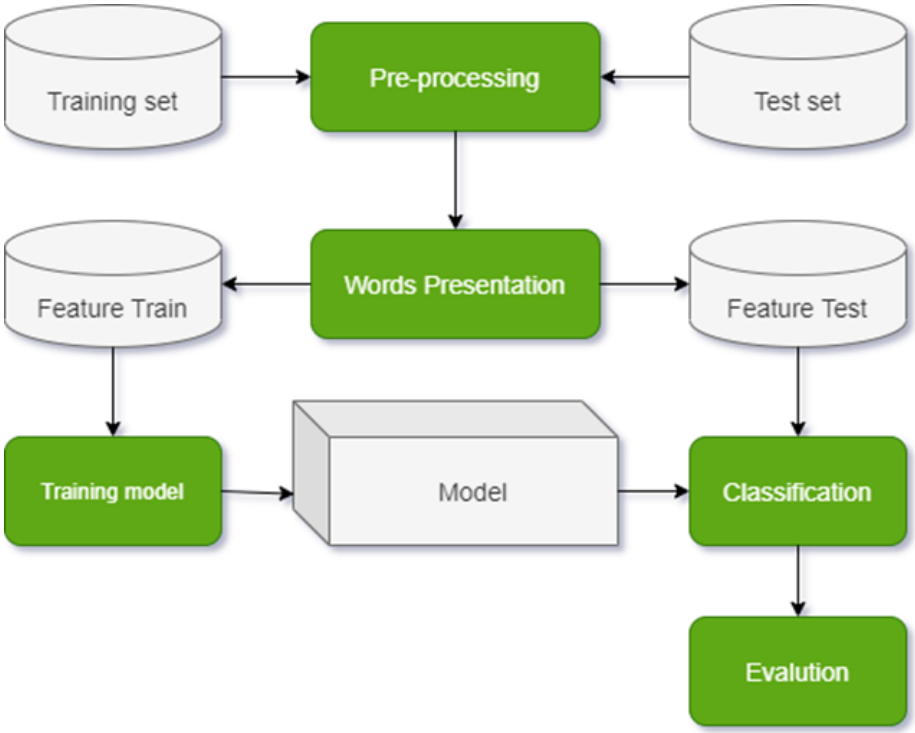
3.1 Nền tảng

Chúng tôi thực hiện bài toán này trên python, Apache Spark và thư viện Spark-NLP. Apache Spark là một framework mã nguồn mở tính toán cụm, Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming). [10]

Spark NLP là một thư viện xử lí ngôn ngữ tự nhiên cho Python, Scala và Java được xây dựng trên Apache Spark ML. Nó cung cập các API cho các PipeLine xử lí với hơn 1100 pipeline trên hơn 192 ngôn ngữ và cũng hỗ trợ các Transformer hiện đại như BERT, XLNet, ELMo. [11]

3.2 Quy trình

Phần này sẽ giới thiệu quy trình chúng tôi đã thực hiện cho bài toán này. Như ở hình 1, dữ liệu chúng tôi sẽ có hai phần là training set và test set, thực hiện các bước tiền xử lí phù hợp trước khi chuyển sang bước biểu diễn từ để trích xuất được các đặc trưng của tranning set và test set. Sau đó các đặc trưng này sẽ được đưa qua mô hình để huấn luyện và sữ dụng các đặc trưng của test set để đánh giá kết quả mô hình.



Hình 1: Tổng quan quy trình

3.3 Tiền xử lí

Chúng tôi thực hiện một số phương pháp tiền xử lí trước khi cho vào mô hình huấn luyện. Các bước thực hiện được mô tả như sau:

- Loại bỏ kí tự đặc biệt
- Loại bỏ các khoảng trống thừa trong câu
- Chuyển toàn bộ câu về chữ thường
- Loại bỏ các từ thường xuất hiện trong câu mà không mang nhiều ý nghĩa khi tính toán - gọi là từ dừng (Ví dụ : I, me, myself,...)

– Chuẩn hoá các từ về dạng chuẩn bằng phương pháp Lemmatization. (Ví dụ: troubled thành trouble, troubling thành trouble)

3.4 Biểu diễn từ và mô hình

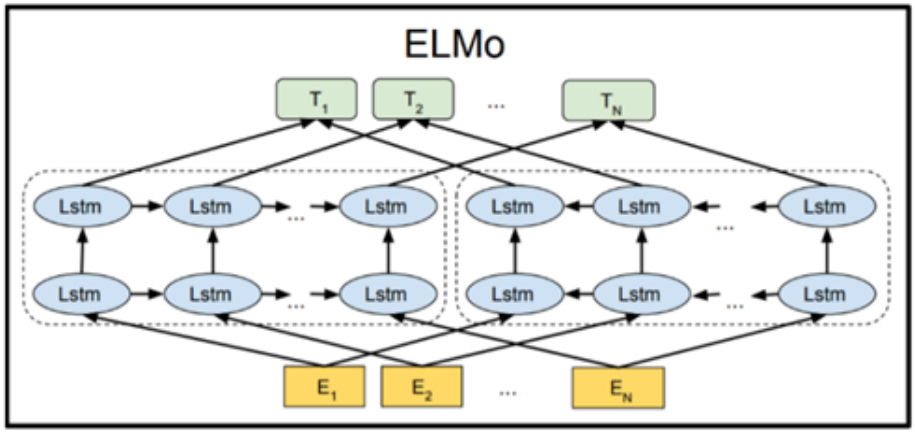
3.4.1 TF-IDF : TF-IDF là phương pháp chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector (VSM), hoặc thành những vector thưa thớt.

3.4.2 GloVe : GloVe (global vectors), một dự án mã nguồn mở của Stanford tạo ra các véc tơ biểu diễn cho các từ. Mô hình là một thuật toán học không giám sát để lấy các biểu diễn vectơ cho các từ. Mô hình ánh xạ các từ vào một không gian vectơ, nơi khoảng cách giữa các từ là sự tương đồng về ngữ nghĩa. Sử dụng các vectơ từ được huấn luyện bởi GloVe, các mô hình có thể tận dụng thông tin về mối quan hệ ngữ nghĩa giữa các từ tốt hơn, từ đó có kết quả tốt hơn trong các bài toán NLP. [4]

3.4.3 ELMO : Thay vì sử dụng cách biểu diễn từ cố định cho từng từ, giống các mô hình như GloVe, ELMo sẽ xem xét toàn bộ câu trước khi chỉ định biểu diễn từ trong đó. Elmo sử dụng LSTM hai chiều trong đào tạo, để mô hình ngôn ngữ của nó không chỉ hiểu từ tiếp theo mà còn cả từ trước đó trong câu. Nó chứa một đường trục LSTM hai chiều 2 lớp. Kết nối dư được thêm vào giữa lớp đầu tiên và lớp thứ hai. Các kết nối dư được sử dụng để cho phép các gradient đi qua mạng một cách trực tiếp, mà không đi qua các chức năng kích hoạt phi tuyến tính. Trực giác cấp cao là các kết nối còn sót lại giúp đào tạo mô hình sâu thành công hơn. Các biểu diễn từ ELMo hoàn toàn dựa trên ký tự, điều này cho phép mạng sử dụng các manh mối hình thái để tạo ra các biểu diễn mạnh mẽ cho các mã thông báo ngoài từ vựng không thể nhìn thấy trong quá trình đào tạo.Không giống như các nhúng từ khác, nó tạo vectơ từ trong thời gian chạy. [5]

3.4.4 BERT : BERT (Bidirectional Encoder Representation from Transformer) là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ (pre-train word embedding). Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng bối cảnh theo cả 2 chiều trái và phải. Cơ chế attention của Transformer sẽ truyền toàn bộ các từ trong câu văn đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional). Đặc điểm này cho phép mô hình học được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó bao gồm cả từ bên trái và từ bên phải. [6]

3.4.5 DistilBERT : DistilBERT học một phiên bản xấp xỉ của BERT, giữ lại 97% hiệu quả dự đoán nhưng chỉ sử dụng một nửa tham số. DistilBERT sử dụng kỹ thuật gọi là distillation, giúp xấp xỉ BERT như một giáo viên của DistilBERT. Ý tưởng ở đây là khi một mạng lớn đã được huấn luyện, phân bố xác xuất đầu ra của nó có thể được xấp xỉ bởi một mạng nhỏ hơn. Hàm loss được sử dụng



Hình 2: Mô hình ELMo

trong xấp xỉ hậu nghiệm trong thống kê Bayes là Kulback Leiber divergence cũng được sử dụng khi huấn luyện DistilBERT. [7]

3.4.6 RoBERTa : RoBERTa được giới thiệu bởi Facebook là một phiên bản được huấn luyện lại của BERT với một phương pháp huấn luyện tốt hơn với dữ liệu được tăng gấp 10 lần. Để tăng cường quá trình huấn luyện, RoBERTa không sử dụng cơ chế dự đoán câu kế tiếp (NSP) từ BERT mà sử dụng kỹ thuật mặt nạ động (dynamic masking), theo đó các token mặt nạ sẽ bị thay đổi trong quá trình huấn luyện. Sử dụng kích thước batch lớn hơn cho thấy hiệu quả tốt hơn khi huấn luyện. RoBERTa sử dụng 160GB văn bản để huấn luyện. Trong đó, 16GB là sách và Wikipedia tiếng Anh được sử dụng trong huấn luyện BERT. Phần còn lại bao gồm CommonCrawl News dataset (63 triệu bản tin, 76 GB), ngữ liệu văn bản Web (38 GB) và Common Crawl Stories (31 GB). [8]

3.4.7 XLM-RoBERTa : XLM-RoBERTa là mô hình RoBERTa đa ngôn ngữ được train trên 100 ngôn ngữ khác nhau. [9]

3.4.8 Logistic Regression : Đây là một trong những phương pháp cơ bản và nổi tiếng của các thuật toán phân loại, đặc biệt là phân loại nhị phân. Trong phân loại văn bản, nó yêu cầu các tính năng thủ công được trích xuất từ dữ liệu. [12]

3.4.9 Support Vector Machine : SVM là thuật toán phổ biến nhất trong học máy cho nhiệm vụ phân loại. Tức là, sau khi đào tạo, chúng ta có được siêu phẳng, nó sẽ phân chia các lớp. Mục đích là tìm một siêu phẳng sao cho khoảng cách giữa đường phân cách và các điểm dữ liệu giữa các lớp là lớn nhất (được gọi là siêu phẳng có lề tối đa). [13]

3.4.10 MultiClassDL : MultiClassDL là một mô hình tích chập sử dụng Bidirectional GRU được xây dựng bằng TensorFlow và hỗ trợ được tối đa 100 nhãn của thư viện Spark NLP

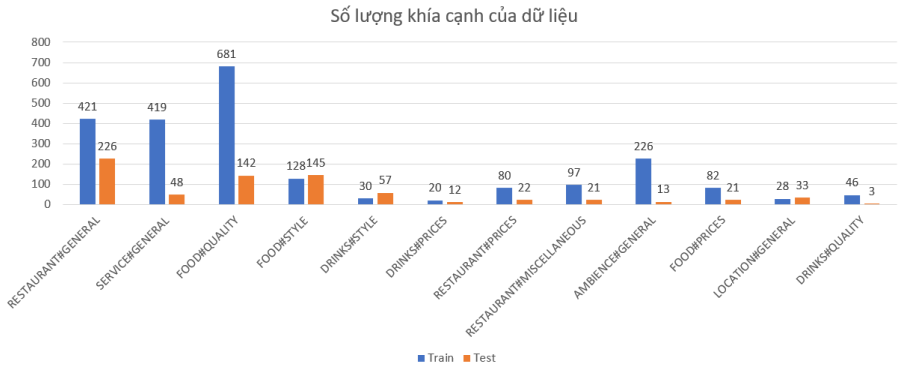
4 Thực nghiệm

4.1 Dữ liệu

Trong bài toán này, chúng tôi thực hiện trên bộ dữ liệu SemEval2016-Task 5 Subtask 1 với miền dữ liệu là nhà hàng ở cấp độ câu [14]. Bộ dữ liệu đã được chia sẵn làm 2 phần là traning set và test set. Ban đầu bộ dữ liệu gồm 2000 câu ở traning set và 676 câu ở test set. Nhưng sau khi thực hiện trích xuất thì chỉ còn 1708 câu ở training set và 587 câu ở test set vì một số câu không được gán nhãn nên chúng tôi đã loại bỏ các câu đó ra khỏi bộ dữ liệu. Các khía cạnh của các câu trong bộ dữ liệu và tổ hợp của 6 thực thể và 5 thuộc tính. Sự kết hợp của các thực thể và thuộc tính này tạo thành 12 khía cạnh cho câu được thể hiện ở hình 3. Bộ dữ liệu có sự chênh lệch ở các khía cạnh là khá lớn, với khía cạnh RESTAURANT#GENERAL, SERVICE#GENARAL, FOOD#QUALITY, AMBIENCE#GENERAL chiếm số lượng vượt trội so với các khía cạnh còn lại, điều này có thể ảnh hưởng đế độ chính xác của mô hình sau này. Chi tiết về sự chênh lêch được thể hiện ở hình 4.

	GENERAL	PRICES	QUALITY	STYLE&OPTION	MISCELLANEOUS
RESTAURANT	√	√	X	X	√
FOOD	X	√	√	√	X
DRINKS	X	√	√	√	X
AMBIENCE	√	X	X	X	X
SERVICE	√	X	X	X	X
LOCATION	√	X	X	X	X

Hình 3: Sự kết hợp giữa các thực thể và thuộc tính



Hình 4: Caption

4.2 Cài đặc thử nghiệm

Trong phần này, chúng tôi trình bày các tham số mà chúng tôi đã thực hiện cho các mô hình của mình với dữ liệu đã qua tiền xử lí. Tuy nhiên để có thể biết được các phương pháp tiền xử lí này có kết quả thế nào, nên chúng tôi cũng thực hiện các mô hình trên cả hai phiên bản là trước và sau tiền xử lí để kiểm tra độ hiệu quả. Với các mô hình, chúng tôi sẽ thực hiện với batch size là 64, epoch là 15, learning rate là 3e-3. Thực hiện đánh giá kết quả mô hình bằng độ đo F1 score trên cả hai thang đo là micro và macro.

5 Kết quả

Chúng tôi đã thực hiện các mô hình theo quy trình với cả hai phiên bản dữ liệu trước và sau tiền xử lí, kết quả của các mô hình học máy và mô hình học sâu trên các bộ biểu diễn từ khác nhau được thể hiện ở bảng 1 và bảng 2.

Bảng 1: Kết quả trên các mô hình học máy

Phương pháp	Không áp dụng tiền xử lí		Có áp dụng tiền xử lí	
	F1-micro	F1-macro	F1-micro	F1-macro
TF-IDF + LogisticRegression	0.4373	0.1403	0.4798	0.1718
TF-IDF + SVM	0.4596	0.1525	0.5019	0.2204

Bảng 2: Kết quả trên các mô hình biểu diễn từ + học sâu

Phương pháp	Không dùng tiền xử lí		Có dùng tiền xử lí	
	F1-micro	F1-macro	F1-micro	F1-macro
GloVe + MultiClassDL	0.6783	0.3729	**0.7160**	**0.5353**
BERT+MultiClassDL	0.7060	0.4973	0.6515	0.4806
UE+MultiClassDL	**0.7374**	**0.5070**	0.7102	0.4696
RoBERTa+MultiClassDL	0.6978	0.5084	0.6395	0.4419
XLM-RoBERTa+MultiClassDL	0.6946	0.4374	0.6149	0.3455
ELMo+MultiClassDL	0.6707	0.4299	0.6917	0.5063
DistilBERT+MultiClassDL	0.7116	0.4682	0.6839	0.4984

Trong bảng 1, ta thấy kết quả các mô hình đều thấp ở cả hai phiên bản dữ liệu, F1-macro rất thấp nên mô hình không hiệu quả đối với bài toán và bộ dữ liệu này. Đối với kết quả học sâu, mô hình áp dụng bộ biểu diễn từ Universal sentence Encoder đạt kết quả cao nhất với F1-score micro là gần 73%, chênh lệch 2% so với phương pháp tốt thứ hai là DistilBERT.

Mô hình trên bộ dữ liệu sau khi xử lí chỉ hiệu quả đối với một số phương pháp cũ như GloVe, SVM, Logistic, các bộ biểu diễn từ hiện đại khi áp dụng tiền xử lí làm giảm kết quả đi nhiều. Các mô hình vẫn còn chênh lệch về độ chính xác ở các nhãn, dẫn đến F1-score macro câu nhất chỉ là 53% của phương pháp GloVe trên bộ dữ liệu sau khi tiền xử lí.

6 Kết luận

Trong báo cáo này, chúng tôi đã áp dụng các phương pháp khác nhau để giải quyết bài toán Aspect Category Detection trên dữ liệu tiếng anh dựa trên nền tảng Apache Spark. Kết quả thu được cao nhất trên độ đo F1-score micro là phương pháp học sâu trên bộ biểu diễn từ Universal sentence Encoder với dữ liệu chưa xử lí (đạt 73%) và trên độ đo F1-score macro là phương pháp học sâu với bộ biểu diễn từ GloVe với dữ liệu đã xử lí (đạt 53%). Các phương pháp tiền xử lí không hiệu quả trên các bộ biểu diễn từ hiện đại và các mô hình vẫn chưa học tốt trên tất cả các nhãn.

Trong tương lai, chúng tôi sẽ tìm cách cải thiện hiệu suất của các mô hình, phát triển các mô hình deep learning khác trên nền tảng Apache Spark và thử áp dụng bài toán này cho tiếng Việt.

Tài liệu

1. X. Zhou, X. Wan, and J. Xiao, “Representation learning for aspect category detection in online reviews,” in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, ser. AAAI’15, Austin, Texas: AAAI Press, 2015, pp. 417–423.
2. W. Xue, W. Zhou, T. Li, and Q. Wang, “MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews,” in Proceedings of the Eighth International Joint Conference on Natural

Language Processing (Volume 2: Short Papers), Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 151–156. [Online]. Available: https://www.aclweb.org/anthology/ I17-2026
3. S. Ramezani, R. Rahimi, and J. Allan, “Aspect category detection in product reviews using contextual representation,” 202
4. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
5. Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).
6. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
7. Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019)
8. Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019)
9. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
10. Spark, Apache. "Apache spark." Retrieved January 17 (2018): 2018.
11. Kocaman, Veysel, and David Talby. "Spark nlp: Natural language understanding at scale." Software Impacts 8 (2021): 100058.
12. Wright, Raymond E. "Logistic regression." (1995).
13. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.
14. Pontiki, Maria, et al. "Semeval-2016 task 5: Aspect based sentiment analysis." International workshop on semantic evaluation. 2016.
15. Zhang, Fangxi, Zhihua Zhang, and Man Lan. "Ecnu: A combination method and multiple features for aspect extraction and sentiment polarity classification." (2014).