Sentiment analysis using Bert

Trần Tuấn Vĩ and Phạm Xuân Thiên

[1] 18520245@gm.uit.edu.vn
[2] 18520158@gm.uit.edu.vn

Tóm tắt nội dung Phân tích cảm xúc là bài toán quan trọng đối với cả nghiên cứu và doanh nghiệp. Hằng ngày, có lượng lớn ý kiến dạng văn bản của người dùng và khách hàng được bày tỏ trên Internet. Trong một thế giới nơi dữ liệu được tạo ra với tốc độ mạnh mẽ như vậy, việc phân tích và xử lý chúng nhanh chóng và chính xác sẽ rất hữu ích. Một trong những công cụ đáp ứng được những tiêu chí này đó là Spark. Trong báo cáo này sẽ thực hiện phân tích cảm xúc và đánh giá trên bộ dữ liệu IMDb Movie Reviews với công cụ SparkNLP và pre-training model là Bert. Đồng thởi thực hiện so sánh với các mô hình SOTA và các thuật toán truyền thống khác.

1 Giới thiệu chung

Sentiment Analysis (Phân tích cảm xúc) là quá trình phân tích, đánh giá quan điểm của một người về một đối tượng nào đó (quan điểm mang tính tiêu cực, tích cực hay bình thường,..). Quá trình này có thể được thực hiện bằng việc sử dụng các tập luật (rule-based), sử dụng Machine Learning (đặc biệt là Deep Learning) hoặc phương pháp Hybrid (kết hợp hai phương pháp trên).
Sentiment Analysis được ứng dụng nhiều trong các sản phẩm thực tế, đặc biệt là trong hoạt động quảng bá kinh doanh. Việc phân tích các đánh giá của người dùng về một sản phẩm xem họ đánh giá tiêu cực, tích cực hoặc đánh giá các mặt hạn chế của sản phẩm sẽ giúp công ty nâng cao chất lượng sản phẩm và tăng cường hình ảnh của công ty. Một ví dụ khác có thể kể đến là việc phân tích quan điểm của người dân về một chính sách, quy định hay dự luật mà nhà nước chuẩn bị ban hành có thể giúp các nhà hoạch định chính sách biết được chính sách nào sẽ mang lại hiệu quả cao và được người dân ủng hộ.
Trong thời gian thực hiện giãn cách vì dịch bệnh, nhu cầu giải trí của người dân ở nhà cần được quan tâm nhiều hơn. Xem phim là hoạt giải trí nhẹ nhàng và hiệu quả nhất lúc này. IMDB[3] là một trang web uy tín lâu năm trong lĩnh vực review phim, có thể cung cấp cho người dùng một góc nhìn chính xác nào đó về bộ phim mà người dùng quan tâm. Để nắm bắt được khách hàng nên IMDB đã phát triển bộ dữ liệu phục vụ cho công việc sentiment analysis về đánh giá phim.
Sentiment analysis đồng thời cũng là một bài toán lớn trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) đã được nghiên cứu trong rất nhiều năm. Vào năm 2018,

[3] https://www.imdb.com/

Google AI đã cho ra mắt một công trình mang tính đột phá trong lĩnh vực xử lý ngôn ngữ tự nhiên là BERT[1] hay Bidirectional Encoder Representations from Transformers.
BERT là một mô hình học sẵn hay còn gọi là pre-train model, học ra các vector đại diện theo ngữ cảnh hai chiều của từ, được sử dụng để ứng dụn vàg các bài toán khác trong lĩnh vực xử lý ngôn ngữ tự nhiên. BERT đã thành công trong việc cải thiện những công việc gần đây trong việc tìm ra đại diện của từ trong không gian số (không gian mà máy tính có thể hiểu được) thông qua ngữ cảnh của nó.
Spark NLP[4] là một thư viện xử lý ngôn ngữ tự nhiên mã nguồn mở, được xây dựng dựa trên Apache Spark và Spark ML. Nó cung cấp một API dễ dàng tích hợp với ML Pipelines và nó được hỗ trợ thương mại bởi John Snow Labs. Thư viện bao gồm nhiều tác vụ NLP phổ biến, bao gồm mã hóa, tạo gốc, lemmatization, một phần của gắn thẻ giọng nói, phân tích cảm xúc, kiểm tra chính tả, nhận dạng thực thể được đặt tên, ...

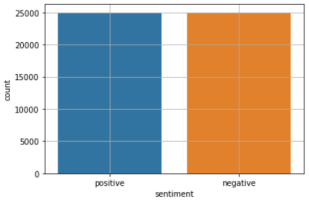
2 Tổng quan bộ dữ liệu

Bộ dữ liệu IMDb Movie Reviews là bộ dữ liệu phân tích cảm xúc. Dữ liệu thu thập từ những đánh giá phim tính theo Internet Movie Database (IMDb). Tổng số câu đánh giá: 50000 câu với 2 nhãn là positive và negative. Câu đánh giá có điểm IMDb < 4 trên 10 thì sẽ được gắn nhãn negative.Và những câu được gắn nhãn positive là những bình luận có IMDb > 7 trên 10 .
Một số đánh giá:
Mỗi phim có không quá 30 câu đánh giá. Số câu mỗi nhãn lần lượt là 25000 và 25000. Số ký tự mỗi câu đánh giá tập trung ở mốc dưới 1000 ký tự. Câu đánh giá tốt dài nhất có 9156 ký tự và 36 ký tự đối với câu ngắn nhất. Còn đối với câu bình luận phê bình thì câu dài nhất là 5653 ký tự. Số từ của câu đánh giá tốt phần lớn nằm trong khoảng dưới 200 từ và đối với đánh giá không tốt thì khu vực tập trung sẽ là dưới 100 từ.
Sau khi chạy một số mô hình máy học cơ bản có một số nhận xét:
Những dữ liệu cần làm sạch: tags, ký tự đặc biệt, url, email. Stopwords không có tác dụng trong quá trình huấn luyện mô hình. Các thông số WordCount và Sentence length không hữu ích Dữ liệu train:test chia theo tỷ lệ 8:2.

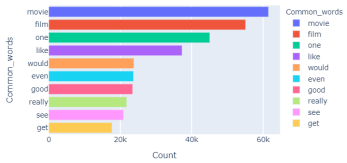
3 Mô hình

Hình 2 mô tả pipeline của mô hình sau quá trình làm sạch dữ liệu.
Giai đoạn đầu tiên là DocumentAssembler() nhận đầu vào là một cột “review” trong Spark data frame (df) tương ứng với cột text trong hình. Sau đó sẽ tạo thêm một cột mới là “document” dạng Document type (AnnotatorType). Có tác dụng tạo ra đúng định dạng đầu vào của Spark NLP.
các “document” được đưa vào SentenceDetector() để tách các câu ra dưới dạng

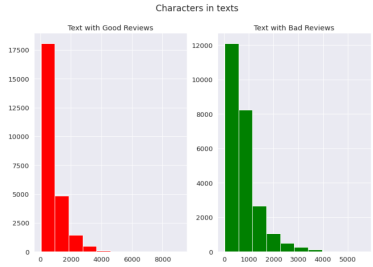
[4] https://nlp.johnsnowlabs.com/



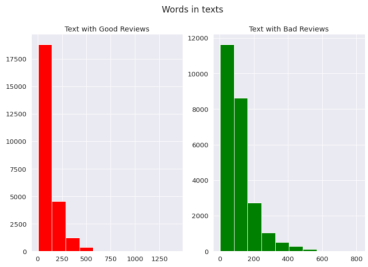
(a) Phân bố dữ liệu



(b) 10 từ thường xuyên xuất hiện

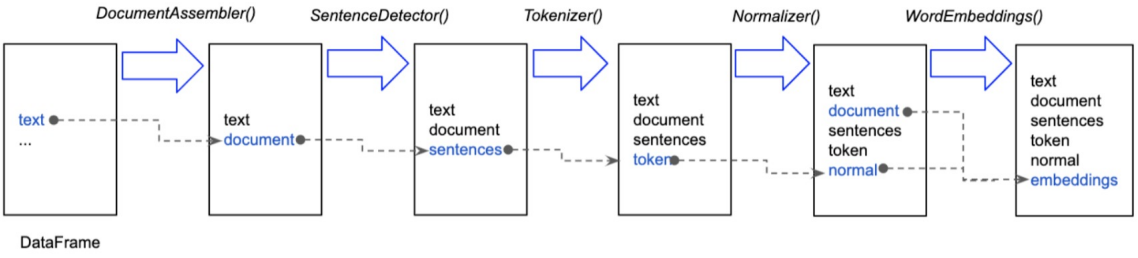


(c) Số lượng ký tự



(d) Phân bố số lượng từ

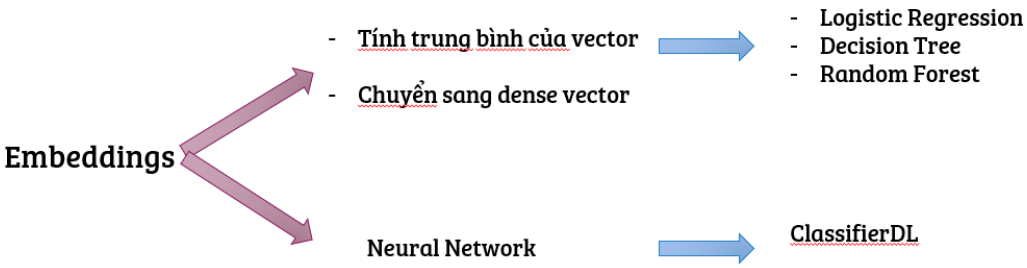
Hình 1: Thống kê



Hình 2: Pipeline

mảng (array) và lưu vào cột mới "sentences". Tiếp theo là giai đoạn tách từ (tokenizer). Mỗi câu sẽ được tách thành các từ và lưu vào cột "token".
Các "document" kết hợp với các token sau khi chuẩn hoá ở giai đoạn Normalizer() - "normal" được sử dụng làm đầu vào của WordEmbedding(). Các phương pháp word embedding được sử dụng là họ Bertology bao gồm: DistilBert[2], Roberta[3] và Albert[4]. Ngoài ra, còn sử dụng thêm Glove embedding và XLnet[5] để so sánh. XLnet là một mô hình SOTA trên bộ dữ liệu IMDB Movie Review được công bố vào năm 2019.

Các vector của word embedding được xử lý theo hai dạng phù hợp với hai dạng thuật toán phân lớp là máy học truyền thống và học sâu. Tổng quan quá trình như hình 3



Hình 3: Các thuật toán phân lớp

4 Kết quả thực nghiệm

Model	P	R	F1
Logistic Regression	0.43	0.43	0.43
Decision Tree	0.44	0.44	0.44
Random Forest	0.45	0.45	0.45
3-Gram + TFIDF	0.49	0.49	0.49
XLnet	0.70	0.70	0.70
Glove	0.83	0.83	0.83
DistilBert	0.82	0.82	0.82
Albert	0.84	0.84	0.84
Roberta	0.84	0.84	0.84

Bảng 1: Kết quả các mô hình trên tập test

Qua kết quả bảng ?? cho thấy, các mô hình Bert cho độ chính xác tối hơn các phương pháp trên Spark. Trong đó Albert và Roberta tốt nhất, nhưng Roberta tối ưu hơn về thời gian training. Để đạt đến 84% thì cần 10 epochs, trong khi Albert cần tới 50 epochs. Bên cạnh đó, các mô hình Bert thế hiện sự ổn định qua các thông số P,R và F1.

5 Kết luận và hướng phát triển

Kết luận:
Bộ dữ liệu đang quá nhỏ. Bert cho kết quả tốt nhất đối với bộ dữ liệu này so với các mô hình thực nghiệm khác. Spark phù hợp với hướng "industry"hơn so với "research". Mô hình SOTA không phù hợp để chạy Spark trong giai đoạn hiện tại.
Hướng phát triển:

Xây dựng thêm bộ dữ liệu để huấn luyện cải thiện độ chính xác mô hình. Tìm hiểu thêm về các hướng xử lý dữ liệu lớn khác. Hướng đến việc streaming dữ liệu bình luận từ các mạng xã hội để đánh giá.

Tài liệu

[1] J. Devlin, M.-W. Chang, K. Lee **and** K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, **jourvol** abs/1810.04805, 2018. arXiv: 1810.04805. **url**: http://arxiv.org/abs/1810.04805.

[2] V. Sanh, L. Debut, J. Chaumond **and** T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, **jourvol** abs/1910.01108, 2019. arXiv: 1910.01108. **url**: http://arxiv.org/abs/1910.01108.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer **and** V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019. arXiv: 1907.11692 [cs.CL].

[4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma **and** R. Soricut, *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2020. arXiv: 1909.11942 [cs.CL].

[5] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov **and** Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *CoRR*, **jourvol** abs/1906.08237, 2019. arXiv: 1906.08237. **url**: http://arxiv.org/abs/1906.08237.