SENTIMENT ANALYSIS BASED ON SENTIMENT LEXICON AND DEEP LEARNING

I. INTRODUCTION

A. TÁC GIẢ:

- Li Yang (Member)
- Ying Li
- Jin Wang (Senior Member)
- R. Simon Sherratt (fellow)

B. TÓM LƯỢC

a. Trong bối cảnh hiện nay, sự bùng nổ về công nghệ kéo theo nền thương mại điện tử phát triển mạnh, Sentiment analysis những user review có thể giúp cải thiện trải nghiệm người dùng.
b. Bài báo này đề xuất một model sentiment analysis mới tên SLCABG (Sentiment Lexicon, Convolutional neural network, Attention-based Bidirectional Gated recurrent unit).
c. Models này sẽ sử dụng ưu điểm của SL và kĩ thuật của Deep Learning.
    i. SL để cải thiện Sentiment features trong reviews.
    ii. CNN và GRU để extract sentiment features chính và context features và attention mechanism để đánh giá trọng số.
    iii. Classify trọng số cho các sentiment features.
d. Bài báo sử dụng data từ trang web dangdang.com, 1 trang web đánh giá sách, data có khoảng 100000 câu reviews.
e. Kết quả của bài toán thể hiện sự vượt trội của model này trong việc cải thiện text sentiment analysis.

C. GIỚI THIỆU

- Với sự phát triển bùng nổ của thương mại điện tử và các nền tảng truyền thông xã hội, càng ngày càng có nhiều user shopping online trên các nền tảng điện tử, shopping online mang lại nhiều tiện ích như: so sánh giá cả, thời gian, đầy đủ mẫu mã, reviews trước đó.
- Việc đánh giá sản phẩm mang lại những đảm bảo về mẫu mã, giá trị sản phẩm, tình trạng sản phẩm qua quá trình, … từ đó giúp người dùng khác biết rõ về sản phẩm đồng thời nhà sản xuất cải thiện trải nghiệm người dùng.
- sentiment analysis hay còn được xem như là bài toán text orientation analysis hay opinion mining.
    - Sentiment lexicon: cốt lõi của sentiment lexicon là xây dựng được từ điển từ vựng với mức độ sentiment, thông qua việc lựa chọn đúng từ tình cảm, trạng từ chỉ mức độ, các từ mang tính âm(đánh giá xấu) hay dương (đánh giá tốt) . Đồng thời xác định cực âm dương và cường độ âm dương cho từ điển từ vựng. sau khi input câu đầu vào, các từ xuất hiện trong từ điển đó sẽ được trích ra, đánh giá và sau đó tổng hợp câu đầu vào để đánh giá giá trị của input, từ đó xác định tính âm dương cho input.
    - Machine learning: phương pháp này sẽ trích xuất các đặc điểm cảm xúc từ dữ liệu manual, vector hóa dữ liệu, sau đó sử dụng machine learning model để phân lớp các đặc điểm các từ đó. Phương pháp này yêu cầu sự can thiệp của con người để thu thập dữ liệu các phân lớp của dữ liệu train đầu vào. Sử dụng các phương pháp học cơ bản

như naives bayes, Support vector machine(SVM), maximum entropy, random forest và conditonal random field model.
    - Deep learning: tiến bộ hơn machine learning khi không cần sự can thiệp của con người, tuy nhiên học sâu yêu cầu dữ liệu lớn hơn hẳn so với machine learning. Deep trích xuất đặc điểm từ nhiều model neural network và học hỏi sai phạm từ chính nó. Model neural network thường được tạo ra từ nhiều hệ thống phân cấp. Một số model thường được sử dụng là CNN, Recurrent NN, Long Short Term Memory (LSTM), Gated Recurrent Unit(GRU).
- Từ những phương pháp trên, bài báo này đề xuất 1 model SLCABG dựa vào sentiment lexicon và deep learning như sau:
    - Đề xuất model dựa vào sentiment lexicon, word vector, CNN, GRU, và attention mechanism, sau đó đánh giá mô hình dựa trên kết quả so sánh trực tiếp trên website.
    - Phân tích dữ liệu đầu vào như: từ đồng nghĩa với các từ cảm xúc, độ dài đoạn văn, số lần lặp lại mô hình và tối ưu hóa mô hình.
-

II. RELATED WORK

A. Sentiment analysis based on sentiment lexicon.

Có thể được cải thiện và tối ưu hóa thông qua nhiều cách khác nhau, tuy nhiên sentiment lexicon lại có một vài hạn chế trong nhiều mảng và chi phí bảo trì thủ công cực kì cao (update review, thêm từ vựng mới qua thời gian, …). Một số cách để cải thiện Sentiment analysis:

- Trích xuất các từ sentiment với chú thích và cường độ của từ.
- Đánh giá đoạn văn cùng với emotes, chọn lọc topic dựa vào từ vựng.
- Phân loại cảm xúc dựa vào đánh dấu label và gắn nhãn yếu để trích xuất các feature hiệu quả.
- Xây dựng mô hình biểu đồ 2 lớp sử dụng emotes và các từ cảm xúc sau đó phân lớp đánh giá model dựa vào các top word trong model đó.

B. Sentiment analysis based on machine learning.

Machine learning có thể tự động trích xuất features, tuy nhiên nó thường phụ thuộc vào manual selection.

Một số phương pháp thực hiện dựa trên Machine learning từng được sử dụng:

- Dựa vào hệ số gini và phân chia bằng SVM.
- Xác định semantic sentiments thông qua một mô hình supervised joint emotion model từ đó xác định sentiment của cả đoạn văn.
- Sử dụng hỗn hợp nhiều machine learning algorithms: naïve bayes, j48, bfTree và OneR.
- Sử dụng SVM và k-nearest neighbors algorithm.
- Sử dụng SVM để phân lớp các post/ comment bằng các mẫu data có sẵn chứa thông tin quan trọng.

C. Sentiment analysis based on deep learning

Deep learning-based không cần thiết sự tham gia của con người để thực hiện, nó có thể tự động hóa lựa chọn và trích xuất tính năng thông qua cấu trúc neural network và học từ chính lỗi của nó.

Một số phương pháp thực hiện dựa trên deep learning từng được sử dụng:

- Sử dụng các features trích xuất từ context và sự lặp lại của các từ được trích xuất để tính toán tính âm dương của đoạn văn.

- Tính toán khoảng cách quan hệ giữa từ target (từ xác định – có thể là từ được trích xuất với giá trị cảm xúc cao) (là từ có thể xác định tính âm dương của đoạn văn) và các từ trong đoạn văn đấy, ví dụ 1 từ mang tính dương + các từ xung quanh có quan hệ cao với từ target sẽ mang lại tính dương cho đoạn văn. Sau đó kết hợp tính toán các từ target trong 1 đoạn văn để có thể xác định âm dương của đoạn văn đó.
- Sử dụng cấu trúc kết hợp giữa CNN và RNN.
- Sử dụng phương pháp divide-conquer, dùng neural network-based model để phân lớp câu, sau đó đặt từng set câu vào CNN để phân lớp sentiment.

III. PHƯƠNG PHÁP

Để cải thiện độ chính xác của sentiment analysis, tác giả kết hợp ưu điểm của lexicon sentiment, CNN model, GRU model, và attention mechanism để tạo ra SCLABG model.

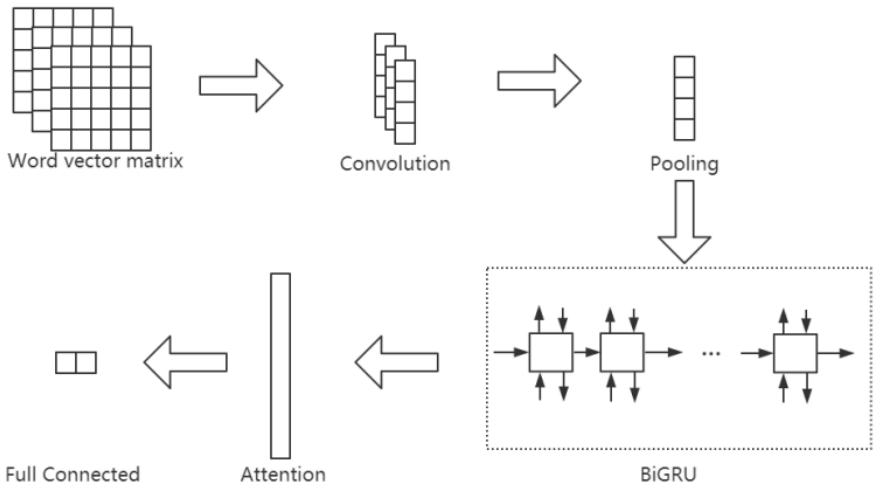
Lexicon sentiment: tăng cường sentiment features trong reviews.

CNN và GRU model: extract main sentiment features và context features trong reviews.

Attention mechanism: tính trọng số cho các features.

⇨ Phân lớp các features kèm với trọng số của nó.

Model hoàn thiện bao gồm 6 phần:



Cấu trúc SLCABG model

A. Constructing an sentiment lexicon

Input: $\mathbf{S = \{ w_1, w_2, \cdots, w_i, \cdots, w_n \}}$

Với S là câu được input, $w_1 \Rightarrow w_n$ là từng từ trong câu.

Sử dụng emotional vocabulary library để tiền xử lý dữ liệu sơ: loại bỏ các từ sentiment mang tính trung lập, đồng thời giữ lại các từ mang tính chất chê hay các từ chê hoặc tính chất khen hay các từ khen.

Các từ này sẽ được chia thành 5 lớp có giá trị 1 3 5 7 9 với mức độ = cường độ của từ. những từ mang tính chất âm sẽ được * với -1 để biến thành cực âm.

w_1 => w_n sau đó sẽ mang giá trị sw hoặc 1 (sw nếu từ có mang tính chất âm dương có thể đánh giá, 1 nếu từ đó không thuộc sentiment lexicon).

Output:

Mảng S với tính chất phần tử như sau:

$$\text{senti}\,(w_i) = \begin{cases} sw_i, & w_i \in SD \\ 1, & w_i \notin SD \end{cases}$$

B. Embedded layer

Input: **S = { w_1,w_2, · · · ,w_i, · · · , w_n}** và $\text{senti}\,(w_i) = \begin{cases} sw_i, & w_i \in SD \\ 1, & w_i \notin SD \end{cases}$

Sử dụng One-hot encoding biến mảng S thành long vector số chiều của vector sẽ bằng với số lượng từ lexicon, mỗi chiều sẽ đảm nhiệm đánh dấu cho 1 từ. đánh dấu 1 cho từ đảm nhiệm nó ở đúng dimension và ở các dimension khác là 0.

Tuy nhiên One-hot vector lại không thể thể hiện quan hệ giữa các từ, bên cạnh đó, số lượng từ trong lexicon là lớn từ đó tạo ra vector word rất lớn và khó sử dụng.

⇨ Encoding word vectors: thể hiện words bằng low-dimensional continuous dense vector (giảm số chiều và tăng độ dày đặc của long vector) và các từ có nghĩa tương đồng sẽ được kết nối tới vị trí chung trong không gian vector. Một số word vector implemetation models thường được dùng gồm : Word2Vec, Glove, ELMo, BERT.

Model BERT là một xét để sử dụng trong bài bởi sự vượt trội về kết quả so với các model khác.

Mỗi w_i trong S được chuyển thành word vector v_i sử dụng BERT model. Với v_i là vector 768 chiều. sau đó, trọng số của vector đó được tính như sau:

$$v'_i = v_i * \text{senti}\,(w_i)$$

Output: vector matrix trọng số **V = { v_1,v_2, · · · ,v_i, · · · , v_n} .**

C. Convolutional layer

Input: **V = { v_1,v_2, · · · ,v_i, · · · , v_n} .**

ở layer này, trích xuất features quan trọng nhất từ matrix.

Thông qua quá trình xử lí convolution hoàn thành dựa vào:

Trọng số weight, chiều cao và rộng của convolution kernel, offset và function ReLU.

Ta sẽ có output vector riêng cho ma trận V' như sau:

Output : **V' = [v_1'', v_2'', … , v_i'', …, v_{n-k+1}'']**

- k là chiều cao của kernel.

D. Pooling layer

Input: vector riêng cho ma trận **V' = [v_1'', v_2'', … , v_i'', …, v_{n-k+1}'']**

Pooling có tác vụ chính là nén text features được đào ra từ convolutional layer sau đó trích xuất các main features. Việc pooling thường được sử dụng qua 2 cách, avarage pooling và max pooling. Đối với text sentiment, thường sẽ chỉ có 1 vài từ mang ảnh hưởng lớn đến tính chất của câu nên ở đây ta sử dụng k-max pooling thay vì 2 phương pháp pooling thường dùng.

Với mỗi phần tử trong V' sẽ được xử lí k-max pooling như sau:

$x = [x_1, x_2, \dots, x_i, \dots, x_{m-k+1}]$

$x_i = \max(v_i''. v_{i+1}'', \dots, v_{i+k-1})$

với m là số chiều của phần tử v_i trong vector V'. và max là function để đặt giá trị cho x_1 trong khoảng từ i đến i+k-1 (k số lượng phần tử - k-max pooling).

Output:

V' = [v_1'', v_2'', … , v_i'', …, v_{n-k+1}'']

Với v_i'' được xử lí qua max pooling

E. BiGRU layer

Input: input matrix (V')

Bidirectional Gated Recurrent Unit là 1 biến thể của mạng RNN truyền thống, với việc set up các gate quyết định thông tin đầu vào và ra sau đó có khả năng đi ngược lại mạng neural network (tính chất bidirectional)

Với h'_t và h''_t là hidden states lấy được khi đi qua mạng neural và đi ngược lại tại thời điểm t thông qua GRU.

Output : tập hợp hidden state tại t:

$$\mathbf{h_t = [h_t';h_t'']}$$

F. Attention layer

Input: $h = [h_1, h_2, \dots, h_n]$

Trong 1 câu review đối với từng từ sẽ có độ quan trọng khác nhau lên phân cực của câu (ví dụ: like, love …) và những từ đối cực của nó (unlike, hate, …). ở attention layer ta sẽ đặt mức độ ảnh hưởng lên các từ trong câu đó.

Trọng số này sẽ được tính toán dựa vào trọng số của ma trận, offset và giá trị của hidden . Sau đó tổng giá trị của các thành phần trong câu:

Output $X = \Sigma_i a_i . h_i$

G. Fully connected layer

Input: feature matrix

Layer này sẽ ánh xạ input qua hàm kích hoạt sigmoid thành một giá trị trong khoảng [0,1]. Càng gần với 0 thì giá trị càng âm, ngược lại càng gần với 1 thì càng dương.

Output: $Y = \text{sigmoid} (W . X) + b$

Với w: weight matrix

b: offset

IV: EXPERIMENTS

A. Dataset

Dữ liệu review sách từ trang web Dangdang, đánh giá 1-5 sao với 1 hoặc 2 sao sẽ là negative, 3,4,5 là positive review.

Thông qua xử lí tay để lọc:

⇨ Dataset bao gồm: 100000 reviews, 50000 negative và positvive.

B. Perfomance metrics

Những thước đo đánh giá bao gồm: precision, recall, F1 score.

TP: số lượng comment phân loại tích cực và bản chất là tích cực (True Positive).

FP: số lượng comment phân loại tiêu cực nhưng bản chất là tích cực (False Positive).

TN: số lượng comment phân loại tiêu cực và bản chất là tiêu cực (True Negative).

FN: số lượng comment phân loại tích cực nhưng bản chất là tiêu cực (False Positive).

Accuracy: tỉ lệ dự đoán đúng

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision: tỉ lệ dự đoán đúng số comment tích cực trên tổng số comment tích cực.

$$P = \frac{TP}{TP+FP}$$

Recall: tỉ lệ dự đoán đúng số comment tích cực trên tổng số comment thực tế.

$$R = \frac{TP}{TP+FN}$$

F1: trọng số trung bình của P và R.

$$F1 = \frac{2*P*R}{P+R}$$

C. DATA PREPROCESSING

- Thực hiện word segmentation. Thêm vào các từ phức mang tính sentiment để đảm bảo các từ đấy biến thành 2 từ riêng lẻ.
- Loại bỏ từ kết câu và các từ không phù hợp với data set.
- Số lượng từ sẽ được đếm đi kèm với số lần xuất hiện và độ dài của từ trong câu. Độ dài trung bình sẽ được đặt ra và trở fixed length (dài hơn sẽ bị chặn, nhỏ hơn giữ nguyên).

Tham số	Giá trị
Độ dài của câu đầu vào	648
Kích thức của word vector	768
Kích thước từ điển	35000
Kích cỡ convolution kernel	3x3
Số hidden neural trong convolution layer	128
Số hidden neural trong BiGRU layer	128
Dropout	0.4

đánh giá bằng cross validation

Method	Accuracy	Precision	Recall	F1
10-fold cross validation	93.2%	92.5%	93.5%	93%
5*2 cross validation	93.3%	93.2%	93.2%	93.2%

Với việc ép fixed length cho câu từ 648 thành 12 thì độ chính xác của 4 độ đo trên giảm khoảng 1.5-1.8%.

Số lượng từ của từ điển từ đồng nghĩa (thesaurus) (20000-50000) đạt độ chính xác tốt nhất ở mức 35000 (precision của 35000 < 20000 và 50000).

Số lần thực hiện cũng tạo ra sự ảnh hưởng lên kết quả và đạt hiệu quả cao nhất ở lần thứ 8, khi qua lần thứ 8 model sẽ bắt đầu bị over fit và kết quả sẽ bị giảm.

Tỉ lệ dropout cũng sẽ đạt giá trị cao nhất ở 0.4, và có sự tăng giảm kết quả ở mốc 0.5 hay 0.2. đồng nghĩa 0.4 là mốc optimize tốt nhất, không phải đạt hiệu quả hoàn hảo nhất.

Việc đánh trọng số cho word vector cũng ảnh hưởng cải thiện lên kết quả 4 độ đo trên:

A: 92.8% => 93.5% P: 92.1% =>93% R: 93.1% => 93.6 % F1: 92.6% => 93.3%

So sánh mô hình SLCABG với các mô hình cơ bản khác thể hiện qua bảng sau.

model	accuracy	precision	recall	F1
NaiveBayes [58]	57.9%	55.6%	79.2%	65.3%
SVM [36]	67.7%	93.8%	38.4%	54.5%
CNN [59]	90.9%	91%	90.2%	90.6%
CNN+Attention [60]	91.4%	90.8%	91.6%	91.2%
BiGRU [61]	92.6%	91.1%	94.1%	92.6%
BiGRU+Attention [62]	93.1%	92.8%	93.2%	93%
SLCABG (Ours)	93.5%	93%	93.6%	93.3%

Có thể thấy mô hình của tác giả cải thiện toàn bộ độ đo so với các mô hình khác (dị biệt BiGRU recall > SLCAB recall, SVM precision > SLCAB precision)

VI. CONCLUSION

Với sự phát triển vượt bậc của e-commerce trong những năm gần đây, sentimentt analysis càng ngày càng nhận được nhiều sự chú ý. Mô hình SLCABG được xây dựng thông qua từ điển từ sentiment, BERT model, CNN model, BiGRU model và attention mechanism. Bằng cách sử dụng model này để phân tích review giúp cải thiện trải nghiệm người dùng cũng như chất lượng của sản phẩm.

Link bài báo : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8970492