Báo cáo thực nghiệm
Họ mô hình ngôn ngữ BERT

Nguyễn Đức Thắng

Giới thiệu bài toán

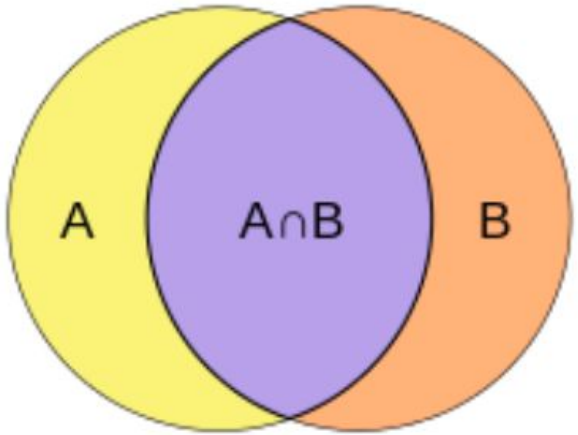
Với tất cả các bình luận lưu hành trên internet, thật khó để có thể biết đằng sau một bình luận, cảm xúc đó sẽ ảnh hưởng như thế nào đến công ty hoặc của một cá nhân nào đó. Vì nó có tính lan truyền. Vì vậy, việc nắm bắt được cảm sắc thái bình luận là việc quan trọng cho mỗi cá nhân và công ty để đưa ra quyết định và xử lý đúng đắn trong từng giai đoạn. Tuy nhiên, vấn đề đặt ra là liệu thực sự cụm từ nào của câu mới đem lại cảm xúc cho toàn câu. Bài toán đặt ra là **trích xuất từ hoặc cụm từ** phản ảnh cảm xúc trong câu.

Dữ liệu

	textID	text	selected_text	sentiment
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative

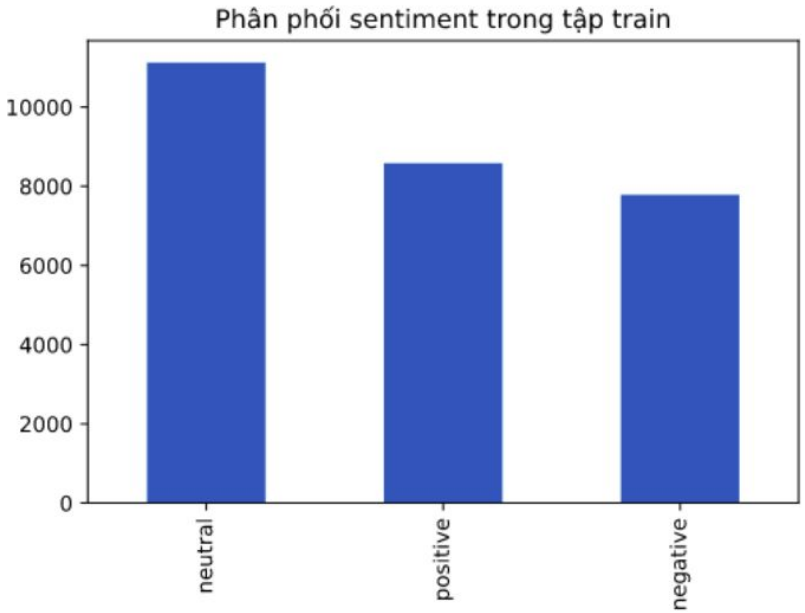
Độ đo đánh giá kết quả

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

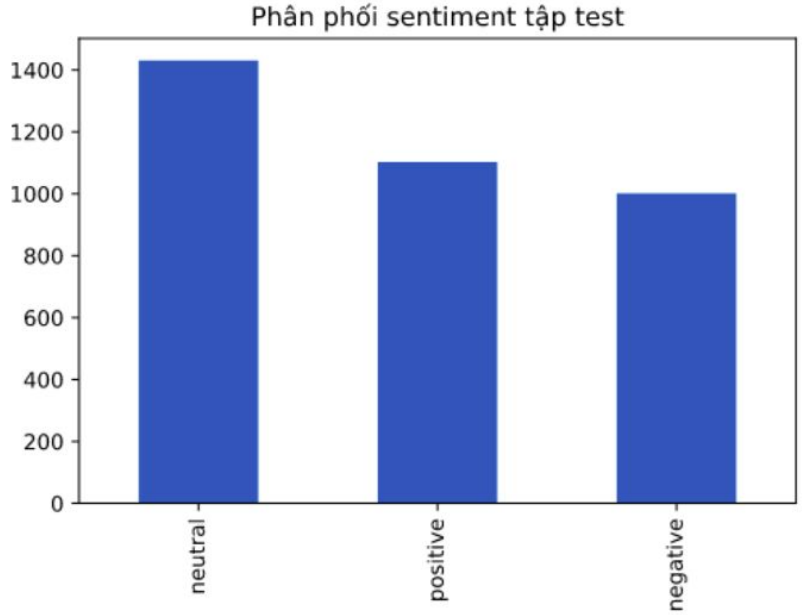


Hình 3.2: Độ đo jaccard

Phân tích dữ liệu



Hình 3.4: Phân phối cảm xúc ở tập huấn luyện

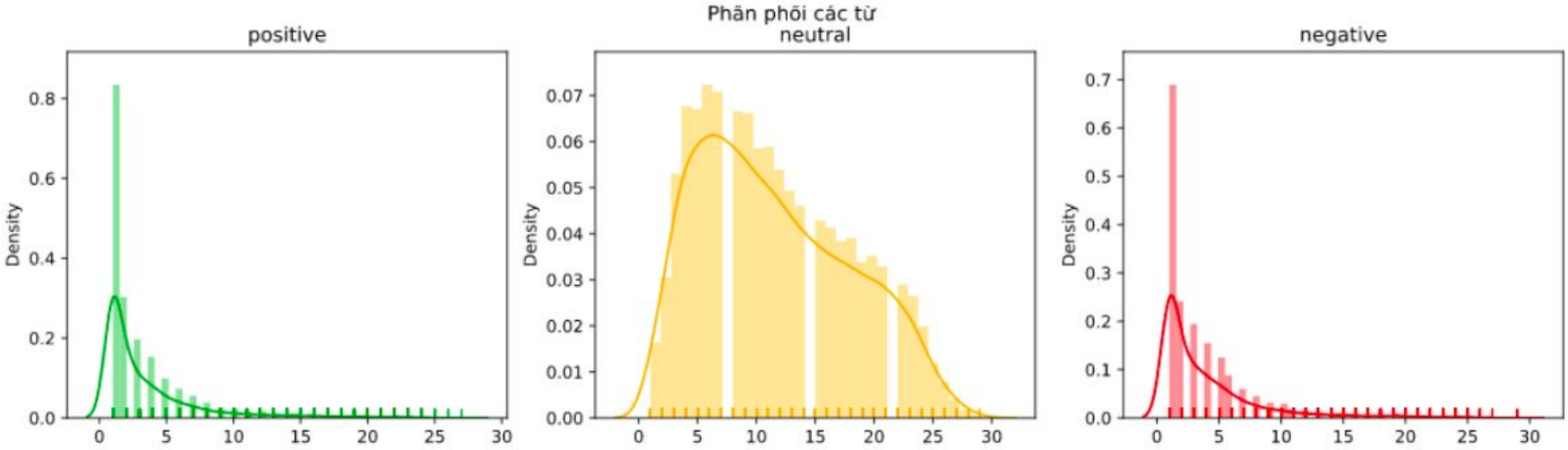


Hình 3.5: Phân phối cảm xúc ở tập kiểm thử

Phân tích dữ liệu

Hình 3.6: Top các từ phổ biến tương ứng với từng trạng thái cảm xúc

Phân tích dữ liệu



Hình 3.7: Phân phối các từ selected text

Thực nghiệm

Thực nghiệm trên 3 mô hình:

1. BERT
2. ROBERTA
3. ALBERT

Biểu diễn dữ liệu đầu vào

tokens	[CLS]	negative	[SEP]	nl	##p	is	very	hard	!	[SEP]	[PAD]	[PAD]
ids	101	4997	102	17953	2361	2003	2200	2524	999	102	0	0
attention_mask	1	1	1	1	1	1	1	1	1	1	0	0
token_type_ids	0	0	0	1	1	1	1	1	1	1	0	0
start_idx	0	0	0	0	0	0	1	0	0	0	0	0
end_idx	0	0	0	0	0	0	0	1	0	0	0	0

Hình 3.8: Biểu diễn dữ liệu đầu vào cho BERT

Biểu diễn dữ liệu đầu vào

tokens	<s>	negative	</s>	</s>	Ġn	lp	Ġis	Ġvery	Ġhard	!	</s>	<pad>	<pad>
ids	0	2430	2	2	295	39031	16	182	543	328	2	1	1
attention_mask	1	1	1	1	1	1	1	1	1	1	1	0	0
token_type_ids	0	0	0	0	1	1	1	1	1	1	1	0	0
start_idx	0	0	0	0	0	0	0	1	0	0	0	0	0
end_idx	0	0	0	0	0	0	0	0	1	0	0	0	0

Hình 3.9: Biểu diễn dữ liệu đầu vào cho RoBerta

Biểu diễn dữ liệu đầu vào

tokens	[CLS]	negative	[SEP]	_	n	lp	_is	_very	_hard	!	[SEP]	<pad>	pad>
ids	2	3682	3	13	103	5478	25	253	552	187	3	0	0
attention_mask	1	1	1	1	1	1	1	1	1	1	1	0	0
token_type_ids	0	0	0	1	1	1	1	1	1	1	1	0	0
start_idx	0	0	0	0	0	0	0	1	0	0	0	0	0
end_idx	0	0	0	0	0	0	0	0	1	0	0	0	0

Hình 3.10: Biểu diễn dữ liệu đầu vào cho ALBert

Pipeline

```
def forward(self, input_ids, attention_mask, token_type_ids):
        # Đầu vào Bert cần chỉ số các token (input_ids)
        # Và attention_mask (Mặt nạ biểu diễn câu 0 = pad, 1 = otherwise)
        # Và token_type_ids
        _, _, hs = self.bert(input_ids, attention_mask, token_type_ids)

        # len(hs) = 13 tensor, mỗi tensor shape là (1, 128, 768)
        x = torch.stack([hs[-1], hs[-2], hs[-3], hs[-4]])
        # x shape (4,1,128,768)
        x = torch.mean(x, 0)
        # x shape (1,128,768)
        x = self.dropout(x)
        x = self.fc(x)
        # x shape (1,128,2)
        start_logits, end_logits = x.split(1, dim=-1)

        # Nếu số chiều cuối là 1 thì bỏ đi (1,128,1) -> (1,128)
        # Ví dụ (AxBxCX1) --> size (AxBxC)
        start_logits = start_logits.squeeze(-1)
        end_logits = end_logits.squeeze(-1)

        return start_logits, end_logits
```

Kết quả thực nghiệm

#	Bert (base)	Roberta (base)	AlBert (base v2)
Jaccard Fold 1	0.6949	0.7068	0.7028
Jaccard Fold 2	0.7026	0.7117	0.7003
Jaccard Fold 3	0.6937	0.7045	0.6881
Jaccard Fold 4	0.6924	0.6922	0.6939
Jaccard Fold 5	0.6950	0.7021	0.7026
Jaccard Mean	0.69572	0.70346	0.69754
Time train/epoch (phút)	33:11	31:41	15:58

Thank you!