So với các nghiên cứu khác có liên quan đến đề tài. Đề tài của chúng tôi có tập đặc trưng được chọn lọc, thử nghiệm nhiều lần và việc đánh trọng số cảm xúc cho bộ từ điển cảm xúc tiếng Việt từ -5 đến +5 sẽ giúp đánh giá giá trị cảm xúc của câu chi tiết hơn là những điều cơ bản giúp nâng cao độ chính xác trong việc phân loại cảm xúc tiếng Việt.

# Từ điển cảm xúc

Chúng tôi xây dựng bộ từ điển cảm xúc tiếng Việt bao gồm năm bộ từ điển nhỏ: từ điển danh từ, từ điển động từ, từ điển tính từ, từ điển trạng từ và từ điển từ tăng cường. Bộ từ điển chủ yếu dựa vào từ điển SO-CAL tiếng Anh. Ngoài ra, chúng tôi bổ sung thủ công một số từ để phù hợp với ngữ pháp tiếng Việt và cách viết ngắn gọn của người dùng mạng xã hội.

Số lượng từ của các bộ từ điển danh từ, động từ, tính từ và trạng từ lần lượt là 1546 từ, 1108 từ, 2357 từ, 749 từ và kèm theo mỗi từ là một số nguyên thể hiện giá trị cảm xúc tương ứng trong phạm vi từ -5 cho hết sức tiêu cực đến +5 cho hết sức tích cực và không có từ nào có giá trị SO là 0.

Bảng 1‑1 Một số từ trong các bộ từ điển danh từ, động từ, tính từ và trạng từ

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Từ điển danh từ | | Từ điển động từ | | Từ điển tính từ | | Từ điển trạng từ | |
| Danh  từ | Giá trị cảm xúc | Động  từ | Giá trị cảm xúc | Tính  từ | Giá trị cảm xúc | Trạng từ | Giá trị cảm xúc |
| hoàn hảo | 5 | tôn kính | 4 | tuyệt vời | 5 | thú vị | 5 |
| lộng lẫy | 4 | hoan hỉ | 4 | cao cấp | 4 | huy hoàng | 4 |
| chiến thắng | 3 | thành công | 3 | bổ ích | 3 | giỏi | 3 |
| phước lành | 2 | sáng tạo | 2 | chặt chẽ | 2 | tươi | 2 |
| độc lập | 1 | tăng | 1 | hợp lý | 1 | sạch | 1 |
| tội phạm | -1 | vùi dập | -1 | cũ | -1 | kỳ quặc | -1 |
| điểm yếu | -2 | xấu hổ | -2 | đần độn | -2 | thô | -2 |

Từ điển từ tăng cường có 185 từ và kèm theo mỗi từ là một số thập phân thể hiện mức độ tăng hoặc giảm giá trị cảm xúc của từ đó.

Bảng 1‑2 Một số từ trong các bộ từ điển từ tăng cường

|  |  |
| --- | --- |
| Từ tăng cường | Mức độ |
| ít | -1.5 |
| chút ít | -0.9 |
| hơi | -0.5 |
| khá | -0.2 |
| chắc | 0.2 |
| siêu | 0.4 |
| hoàn toàn | 0.5 |

Ví dụ: Từ “nhếch nhác” có giá trị cảm xúc là (-3) thì “khá nhếch nhác” có giá trị cảm xúc là (-3)\*(1 – 0,1) = (-2.7). Từ “xuất sắc” có giá trị cảm xúc là 5 thì “xuất sắc nhất” có giá trị cảm xúc là 5\*(1 + 1) = 10.

# Bộ dữ liệu huấn luyện

Để xây dựng bộ dữ liệu huấn luyện. Chúng tôi tiến hành gán nhãn cho câu. Đây là quá trình mang tính chủ quan của từng cá nhân. Cảm xúc của con người rất phức tạp. Vấn đề đầu tiên là phân loại xem câu có chứa cảm xúc hay không chứa cảm xúc (thường được gọi là phân loại chủ quan). Vấn đề thứ hai là phân loại những câu chứa cảm xúc là tích cực hay tiêu cực (thường được gọi là phân loại cảm xúc).

Chúng tôi quyết định lựa chọn 3 bộ dữ liệu ở 3 chủ đề: giáo dục, phim ảnh và thể thao. Mỗi bộ dữ liệu bao gồm từ hơn 250 cho 350 bình luận về các chủ đề trên. Sau đó, từ 03 bộ dữ liệu trên chúng tôi gộp tất cả các chủ đề lại để xây dựng một bộ dữ liệu huấn luyện lớn hơn gồm 885 câu. Đây là bộ dữ liệu tổng hợp.

Đầu tiên chúng tôi thực hiện phân loại chủ quan bằng tay với các bộ dữ liệu trên. Kết quả phân loại chủ quan bằng tay được thể hiện trong bảng sau:

Bảng 2‑1 Kết quả phân loại chủ quan bằng tay

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Chủ đề | Dữ liệu huấn luyện | |
| Câu chủ quan | Câu khách quan |
| 1 | Giáo dục | 173 | 99 |
| 2 | Phim ảnh | 194 | 95 |
| 3 | Thể thao | 248 | 76 |
| 4 | Tổng hợp | 615 | 270 |

Sau khi phân loại chủ quan, chúng tôi lựa chọn những câu chủ quan (có cảm xúc) để tiếp tục phân loại cảm xúc. Kết quả phân loại được trình bày theo bảng sau:

Bảng 2‑2 Kết quả phân loại cảm xúc bằng tay

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Chủ đề | Dữ liệu huấn luyện | |
| Câu tích cực | Câu tiêu cực |
| 1 | Giáo dục | 133 | 40 |
| 2 | Phim ảnh | 115 | 79 |
| 3 | Thể thao | 201 | 47 |
| 4 | Tổng hợp | 449 | 166 |

# Mô hình thực nghiệm

## Phương pháp phân loại chủ quan

Để đánh giá một câu có hay không có cảm xúc chúng tôi lựa chọn những đặc trưng sau:

* Đặc trưng số 1: số lượng từ trong câu. Số lượng từ trong câu cũng thể hiện cảm xúc mà người nói, người viết muốn biểu lộ với người nghe, người đọc. Nếu số lượng từ lớn thông thường đó sẽ là một câu có cảm xúc vì người nói, người viết đã đầu tư một công sức đáng kể và rõ ràng là họ quan tâm đến chủ đề đang được nhắc đến.
* Đặc trưng số 2, 3, 4 và 5: tổng giá trị cảm xúc của các từ loại: tính từ, trạng từ, danh từ và động từ trong câu. Giá trị cảm xúc trong câu phụ thuộc vào loại từ và giá trị cảm xúc của loại từ đó được so khớp với bộ từ điển SO-CAL tiếng Việt. Chúng tôi nhận thấy, giá trị cảm xúc trong câu chủ yếu phụ thuộc vào các loại từ sau: trạng từ, tính từ, danh từ và động từ.
* Đặc trưng số 6: tổng giá trị cảm xúc của câu. Đặc trưng này thể hiện tổng giá trị cảm xúc của câu. Giá trị của đặc trưng này về cơ bản là tổng của 04 đặc trưng phía trên mà chúng tôi xây dựng. Mặc dù chúng có liên quan với nhau và tưởng chừng giá trị này dư thừa, nhưng thực tế việc tính tổng này là hết sức cần thiết vì nếu tổng những giá trị phía trên bằng 0 thì việc đánh giá chủ quan còn chưa chắc chắn là đúng đắn.

Ngoài ra, giá trị cảm xúc trong một câu không chỉ phụ thuộc vào từ hàm chứa cảm xúc, một câu chủ quan còn phụ thuộc vào loại câu của nó nữa. Nếu là một câu nghi vấn hoặc một câu điều kiện thì câu đó hoàn toàn không có giá trị cảm xúc. Do đó, tổng giá trị cảm xúc của câu còn có thể bằng 0 nếu như câu đó thuộc một trong hai loại câu bên trên.

## Phương pháp phân loại cảm xúc

Sau khi phân loại những câu chủ quan có hàm chứa cảm xúc, chương trình sẽ tiếp tục việc phân loại cảm xúc cho những câu này. Việc phân loại cảm xúc của một câu thực tế là việc lựa chọn bộ đặc trưng tốt để đạt được độ chính xác cao. Bộ đặc trưng sau đây chúng tôi lựa chọn được kế thừa từ phương pháp phân tích cảm xúc ở tiếng Anh đồng thời có sự phát triển và chỉnh sửa cho phù hợp với đặc trưng ngôn ngữ tiếng Việt.

* Giá trị cảm xúc của câu phụ thuộc vào từ hàm chứa cảm xúc: Đầu tiên, những đặc trưng cơ bản nhất là sự kế thừa từ phương pháp phân tích chủ quan. Bao gồm: Giá trị cảm xúc của các loại từ trong câu: tính từ, trạng từ, danh từ và động từ. Tổng giá trị cảm xúc của tất cả các loại từ bên trên.
* Giá trị cảm xúc của câu phụ thuộc vào từ tăng cường: giá trị cảm xúc của từ chịu ảnh hưởng của từ tăng cường được tính theo công thức:

*Giá trị cảm xúc = giá trị từ tăng cường \* giá trị cảm xúc của từ*

Tổng những giá trị này sẽ là giá trị cảm xúc mới của câu sau khi xét từ tăng cường. Trong trường hợp không có từ tăng cường trong câu, giá trị này chính là giá trị của tổng giá trị cảm xúc của tất cả các loại từ trong câu.

* Giá trị cảm xúc của câu phụ thuộc vào từ phủ định: danh sách từ phủ định bao gồm: “không”, “không có”, “không phải”, “không được”, “chẳng”, “chẳng có” và “chẳng phải”. Những từ có cảm xúc chịu ảnh hưởng của từ phủ định trong câu được xác định lại giá trị cảm xúc của nó theo công thức sau:

*Giá trị cảm xúc = (-1) \* giá trị cảm xúc của từ*

* Giá trị cảm xúc của câu phụ thuộc vào từ khiếm khuyết: nếu câu có từ khiếm khuyết (“nên”, “phải”, “có thể”, v.v…) thì giá trị cảm xúc của cả câu được định nghĩa lại theo công thức:

*Giá trị cảm xúc của câu = (0.5) \* tổng giá trị cảm xúc các loại từ trong câu*

* Giá trị cảm xúc của câu có xu hướng tích cực: trong thực tế và văn hóa Việt Nam. Việc sử dụng từ ngữ nói giảm, nói tránh để thể hiện cảm xúc là hết sức phổ biến. Người dùng thường tránh nói ra những từ ngữ tiêu cực, do đó dẫn đến việc những từ ngữ tiêu cực thường ít gặp hơn so với từ tích cực. Theo đó, những từ ngữ hàm chứa cảm xúc tiêu cực (mang giá trị cảm xúc âm) sẽ được tính theo công thức:

*Giá trị cảm xúc = (1 + 0.5) \* giá trị cảm xúc của từ*

* Giá trị cảm xúc của câu có từ liên kết mang ý nghĩa trái ngược: danh sách từ liên kết mang nghĩa trái ngược: “nhưng”, “nhưng mà”, “mà” và “cơ mà”. Đối với những câu có chứa những từ liên kết mang ý nghĩa trái ngược được nêu bên trên thì giá trị cảm xúc của câu không phải là giá trị của tổng giá trị cảm xúc các loại từ trong câu mà chỉ là giá trị cảm xúc của về phía sau từ liên kết đó. Do đó, giá trị cảm xúc của câu được xác định theo công thức:

*Giá trị cảm xúc câu = tổng giá trị cảm xúc – tổng giá trị cảm xúc đứng trước từ liên kết*

# Kết quả thực nghiệm

## Kết quả đánh giá phương pháp phân loại chủ quan

Từ bộ dữ liệu thử nghiệm phân loại bằng tay, phương pháp phân lớp SVM và bộ dữ liệu huấn luyên được trình bày bên trên chúng tôi tiến hành kiểm tra độ chính xác của phương pháp phân loại chủ quan. Kết quả đánh giá mức độ chính xác theo bảng sau:

Bảng 4‑1 Kết quả đánh giá độ chính xác phương pháp phân loại chủ quan

|  |  |  |
| --- | --- | --- |
| STT | Chủ đề | Kết quả thử nghiệm (độ chính xác: %) |
| 1 | Giáo dục | 92.6% |
| 2 | Phim ảnh | 89.7% |
| 3 | Thể thao | 89.5% |
| 4 | Tổng hợp | 89.8% |

## Kết quả đánh giá phương pháp phân loại cảm xúc

Chúng tôi tiếp tục tiến hành đánh giá độ chính xác của phương pháp phân loại cảm xúc. Kết quả được trình bày trong bảng sau:

Bảng 4‑2 Kết quả đánh giá độ chính xác phương pháp phân loại cảm xúc

|  |  |  |
| --- | --- | --- |
| STT | Chủ đề | Kết quả thử nghiệm (độ chính xác: %) |
| 1 | Giáo dục | 90.8% |
| 2 | Phim ảnh | 79.2% |
| 3 | Thể thao | 95.0% |
| 4 | Tổng hợp | 89.5% |