**CỘNG HÒA XẪ HỘI CHỦ NGHĨA VIỆT NAM  
Độc lập – Tự do – Hạnh phúc  
-------------------------------**



Báo cáo tìm hiểu   
phương pháp phân tích quan điểm  
dựa trên từ vựng (Lexicon-based approach)

Hà Nội – 2022

Mục lục

[I. Khái quát về phân tích quan điểm 4](#_Toc108682160)

[1.1. Định nghĩa 4](#_Toc108682161)

[II. Phân tích quan điểm dựa trên từ vựng 4](#_Toc108682162)

[2.1. Khái quát 4](#_Toc108682163)

[2.2. Các bước cơ bản của hướng tiếp cận 4](#_Toc108682164)

[2.3. Các bộ từ vựng và mô hình từ vựng thường được sử dụng 4](#_Toc108682165)

# Khái quát về phân tích quan điểm

## Định nghĩa

Phân tích quan điểm là phương pháp bóc tách *quan điểm chủ quan* và *tính phân cực* từ một đoạn văn bản. Thông qua tính phân cực và độ mạnh của các từ/ cụm từ, phương pháp sẽ cho ra kết quả là quan điểm của nội dung đầu vào là tích cực, tiêu cực hay trung tính.

* Quan điểm chủ quan(subjective): Thường chứa các nội dung mang tính quan điểm cá nhân người viết/người nói, như một dự đoán hay kết luận.
* Tính phân cực(polarity): Biểu thị tính tích cực/ tiêu cực của đoạn văn bản bằng một chỉ số thập phân trong khoảng [-1,1], cụ thể:

+ Quan điểm tiêu cực: Tính phân cực < 0

+ Quan điểm trung tính: Tính phân cực ~0

+ Quan điểm tích cực: Tính phân cực > 0

Ngoài phân cực theo quan điểm (tích cực – tiêu cực), nội dung của văn bản có thể mang tính chủ quan (objective - tức chứa các sự việc, sự thật nhiều hơn là quan điểm). Nội dung như vậy sẽ không mang tính tiêu cực hay tích cực, do đó các câu này sẽ không thể đem ra phân tích quan điểm.

…

Có nhiều hướng tiếp cận để phân tích quan điểm, báo cáo này sẽ tập trung vào hướng tiếp cận dựa trên từ vựng (Lexicon-based Aproach)

# Phân tích quan điểm dựa trên từ vựng

## Khái quát

Trước khi mua một sản phẩm hay sử dụng một dịch vụ, mọi người thường có xu hướng tìm kiếm và tham khảo các nhận xét, đánh giá trên mạng về dịch vụ/ sản phẩm đó. Trong các đánh giá này thường chứa những từ/cụm từ mang tính biểu đạt cảm xúc như *“tuyệt”* (tích cực) hay *“tồi”* (tiêu cực). Qua những từ như thế có thể đem tới cho người đọc những ấn tượng đầu tiên về sản phẩm đó.

Trong phương pháp phân tích quan điểm dựa trên từ vựng, các từ sau khi được tách ra từ câu và xử lí sẽ được đánh nhãn *tích cực* hoặc *tiêu cực* (đôi khi sẽ là trung tính) bằng cách ánh xạ các từ này qua một bộ “từ điển giá trị hóa”.

Sau khi các từ đã được gắn nhãn, chúng ta có thể dùng các phép tính tổng hoặc tính trung bình để suy ra một điểm số tổng thể cho cả một nội dung văn bản. Một trong những công thức thường dùng nhất là công thức tính giá trị đánh giá quan điểm (StSc: sentiment score):

StSc =

Nếu điểu StSc là âm, đoạn văn bản sẽ được phân loại là tiêu cực. Tiếp theo đó nếu điểm StSc là dương thì văn bản là tích cực, StSc là 0 thì văn bản sẽ là trung tính.

*Lưu ý:* Cách tiếp cận dựa trên từ vựng này không dùng kĩ thuật học máy. Kết quả đánh giá quan điểm sẽ được tính toán và cho ra ngay lập tức tùy theo bộ từ vựng được sử dụng để đánh nhãn từ.

## Các bước cơ bản của hướng tiếp cận

### Tiền xử lý

Trước khi có thể gắn nhãn và tính điểm, đánh giá, dữ liệu văn bản đầu vào cần được đi qua các bước làm sạch và xử lý cơ bản. Có thể kể đến như:

* Loại bỏ các dấu chấm câu và kí tự đặc biệt
* Loại bỏ các từ được dùng nhiều và ít giá trị khi đánh giá quan điểm như: à, ừ, bị, các, cái… Hiện nay để hỗ trợ cho việc loại đi các từ này đã có sẵn một vài repository trên github đã được đóng góp và xây dựng từ trước nên không cần thực hiện lại việc tổng hợp các từ này.

<Tham khảo: https://github.com/stopwords/vietnamese-stopwords>

* Đưa các từ thường hay viết tắt về dạng đầy đủ (ko -> không, uk -> ừ, HN -> Hà Nội,…)
* Đưa các câu văn thành từng từ nhỏ (tokenization)

## Các dữ liệu từ vựng thường được sử dụng

Đã có nhiều bộ từ vựng được xây dựng sẵn nhằm thực hiện được cách tiếp cận theo từ vựng, nhưng đa số là tiếng Anh, và vẫn còn tương đối ít bộ từ vựng sẵn được xây dựng bằng tiếng Việt.

### Các bộ từ tiếng Anh

* **AFINN:** Là bộ từ vựng đơn giản và phổ biến nhất trong phân tích quan điểm, chứa 3382 từ cùng với điểm phân cực của từng từ.

**<**Xem thêm: <https://github.com/fnielsen/afinn> >

* **SentiWordNet:** Là bộ từ vựng dành cho khai thác quan điểm, mỗi từ trong bộ từ sẽ bao gồm 3 chỉ số thể hiện quan điểm: tích cực, tiêu cực và khách quan

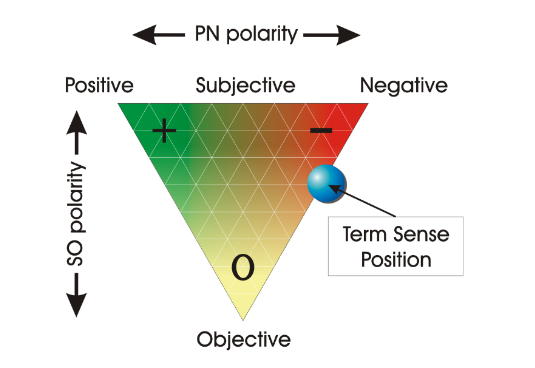


Figure II‑1: Hình minh họa cho 3 chỉ số của SentiWordNet

* **VADER (Valence Aware Dictionary and Sentiment Reasoner):** Là một bộ từ vựng và còn là một công cụ thông minh trong phân tích quan điểm. Công cụ này thậm chí có thể phát hiện tính phân cực quan điểm trong câu văn cũng như nhận biết và điều chỉnh được “mức độ cảm xúc” của một câu văn chưa hề được gắn nhãn.

### Các dữ liệu từ tiếng Việt

Hiện nay các bộ dữ liệu tiếng Việt đều đang được tích cực xây dựng và phát triển. Hầu hết các bộ tự vựng tiếng Việt đã được sử dụng đều nằm trong các nghiên cứu, đặc biệt là từ các nghiên cứu và chiến dịch được tổ chức thường xuyên của VLSP (Vietnamese Language and Speech Processing). Chiến dịch đầu tiên của Việt Nam về phân tích quan điểm được tổ chức tại VLSP 2016 – với đề bài tập trung chủ yếu vào phân loại phân cực (tích cực, tiêu cực, trung tính). Các dữ liệu này hiện có thể truy cập thông qua trang web lưu trữ của VLSP: <https://vlsp.org.vn/resources>

Một nghiên cứu tham khảo:

[1]: Huyen T M Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, Cuong A Le, VLSP Shared Task: Sentiment Analysis, Journal of Computer Science and Cybernetics, Vol 34, No 4, pp. 283-294, 2018.

[2]: Huyen T M Nguyen, Quyen T Ngo, Luong X Vu, Vu M Tran, Hien T T Nguyen, VLSP Shared Task: Named Entity Recognition, Journal of Computer Science and Cybernetics, Vol 34, No 4, pp. 295-310, 2018.

[3]: Xuan-Son Vu , Thanh Vu , Mai-Vu Tran , Thanh Le-Cong, Huyen T M. Nguyen: HSD Shared Task in VLSP Campaign 2019: Hate Speech Detection for Social Good