Creating Domain-Specific Lexicon

Phân tích cảm xúc nhắm tới xác định, phân loại quan điểm của khách hàng

đánh giá dựa vào hướng tiếp cận học có giám sát và không giám sát

được thực hiện bởi:

- công nghệ học máy như Bayes, SVM

- thuật ngữ cảm xúc (map đến 1 điểm số đánh giá)

Xây dựng từ điển miền cụ thể dựa vào xác suất và các kỹ thuật lý thuyết thông tin

Kỹ thuật khai phá hoạt động tốt như các cách tiếp cận truyền thống

và từ vựng miền tốt hơn so với từ chung

Phân tích cảm xúc chia thành 2 loại:

- dựa vào trên ngữ liệu

xây dựng phân loại từ đối tượng đã gán nhãn (ML phân loại có giám sát)

- dựa vào từ vựng

học không giám sát, sử dụng từ điển các thuật ngữ với phân loại P/N có trọng số

Phân cực rời rạc: tích cực, tiêu cực, trung tính

(MPQA Subjectivity lexicon)

Phân cực liên tục: +1 đến -1

Phân cực thông qua cảm xúc: ứng với 1 giá trị rời rạc trong danh sách

(word-emotion association lexicon)

Phân cực phân số: 3 số dương tổng bằng 1 mỗi giá trị ứng vs P, Neg, Neu

(SentiWordNet lexicon)

Lợi ích của hướng tiếp cận lexicon:

- lexicon được xây dựng từ kho ngữ liệu lớn có thể sử dụng trong các ứng dụng khác, nơi có thể không đủ thông tin để thực hiện các pp tiếp cận dựa trên ngữ liệu (ML)

- được sử dụng rộng rãi cho phân loại cảm xúc cho đa ngôn ngữ

- phân loại cảm xúc theo miền

Phân tích cảm xúc nhạy cảm với miền áp dụng vì từ có thể mang ngữ nghĩa khác nhau đối với miền khác nhau

1 từ không có duy nhất 1 phân loại đối với tất cả các miền

các hướng tiếp cận hiện tại chủ yếu nghiên cứu thích ứng, chuyển đổi cảm xúc từ lớp phân loại đã được huấn luyện hoặc thuật ngữ từ 1 miền tới miền khác bắt đầu từ thuật ngữ chung

=> rất ít cách làm mà không yêu cầu kiến thức sẵn.

=> Cách làm:

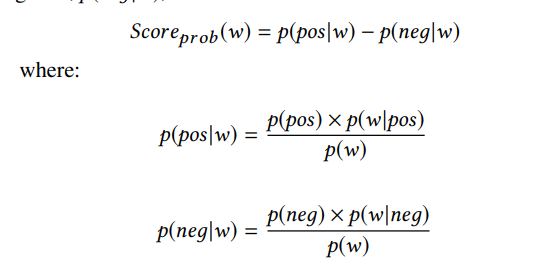
tính điểm cảm xúc miền riêng biệt dựa vào khai phá văn bản

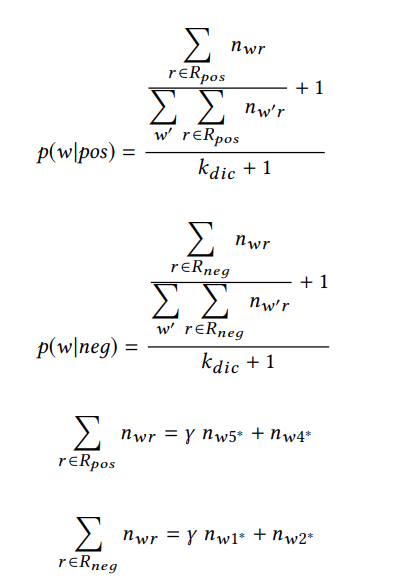
Ước tính điểm của từ:

kết hợp điểm xác suất và điểm lý thuyết thông tin

- điểm xác suất

tính bằng độ lệch khả năng là từ tích cực và khả năng là từ tiêu cực

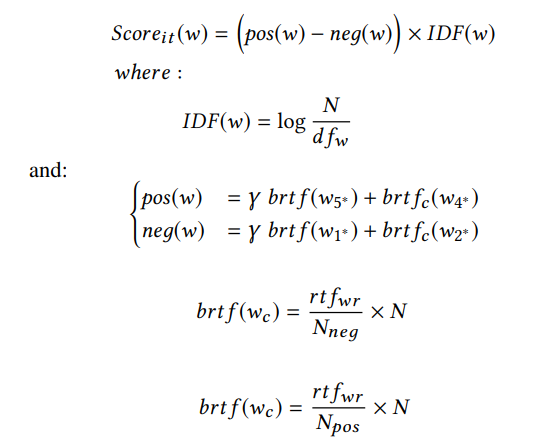




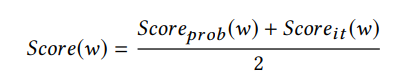
- điểm lý thuyết thông tin

Dựa vào TF-IDF đánh giá độ quan trọng của 1 từ trong văn bản

Tính bằng độ lệch điểm tích cực và tiêu cực nhân với nghịch đảo tần số



=> Điểm cuối cùng:



Improve Opinion Aspect Extraction

Từ tập opinion xây dựng sẵn => tập aspects dựa vào quy tắc cú pháp (DP)

=> Chưa đủ tốt

cải thiện: (1) tương đồng ngữ nghĩa

(2) quan hệ liên kết

(1): giải quyết vấn đề thiếu aspects đồng nghĩa của DP => sử dụng Word Vectors

(2): aspects cùng miền => sử dụng luật liên kết

Thuật toán:

AER (Aspect Extraction based on Recommendation)

- cho tập document và tập seed words chỉ quan điểm

dùng tập luật DP precision cao => tập trích xuất có precison cao nhưng recall k cao : T-

dùng tập luật DP recall cao => tập trích xuất có recall cao nhưng precision k cao : T+

T = (T+ - T-)

Similarity-based Recommendation

Ts = Sim-recom(T-, T)

Ts = []

for t in T:

if Sim(t,T-) >= e # hàm tính độ tương đồng dựa vào Word Vectors (dùng gensim với pre-trained model)

Ts.append(t)

(có thể bỏ qua)

Association-based Recommendation

Ta = AR-recom(T-, T)

=> Tập aspects T- &Ts & Ta