XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG 1 – PHÂN TÍCH Ý KIẾN

NGUYỄN TRỌNG CHỈNH

TRÌNH BÀY

1. GIỚI THIỆU
2. MÔ HÌNH NAÏVE BAYES
3. MÔ HÌNH LOGISTIC REGRESSION
4. ĐÁNH GIÁ

1. GIỚI THIỆU

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Khái niệm phân tích ý kiến:

Phân tích ý kiến (Sentiment Analysis – SA, Opinion Mining) là phân tích một đoạn văn bản để xác định thái độ đối với sự vật, con người, hay sự kiện được diễn tả trong đoạn văn bản đó là tích cực (positive), tiêu cực (negative) hay trung tính (neutral).

Ví dụ:

- Bàn phím này **rất cứng**! $\Rightarrow$ negative
- Các khe cắm PCIE trên mainboard **rất cứng**! $\Rightarrow$ positive

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Vì sao cần phân tích ý kiến:

Amazon customer reviews

Customer reviews

4.7 out of 5

8,068 global ratings

5 star	84%
4 star	9%
3 star	3%
2 star	1%
1 star	3%

How customer reviews and ratings work

Customers say

Customers find this DSLR camera to be a great entry-level option that's easy to use and offers good value for money, with one customer noting it costs less than a single headshot session. The camera receives positive feedback for its picture quality, with one review mentioning exceptionally clear results in well-lit areas, and its features, including preset modes and a real-time feature guide. The functionality and autofocus performance receive mixed reviews, with some customers praising the auto focus while others report issues with the camera not working properly.

Generated from the text of customer reviews

Select to learn more

✓ Camera quality | ✓ Picture quality | ✓ Ease of use | ✓ Value for money | ✓ Camera for beginners | ✓ Features | Functionality | Autofocus

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Vì sao cần phân tích ý kiến:

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 122-129. DOI:10.1609/icwsm.v4i1.14031

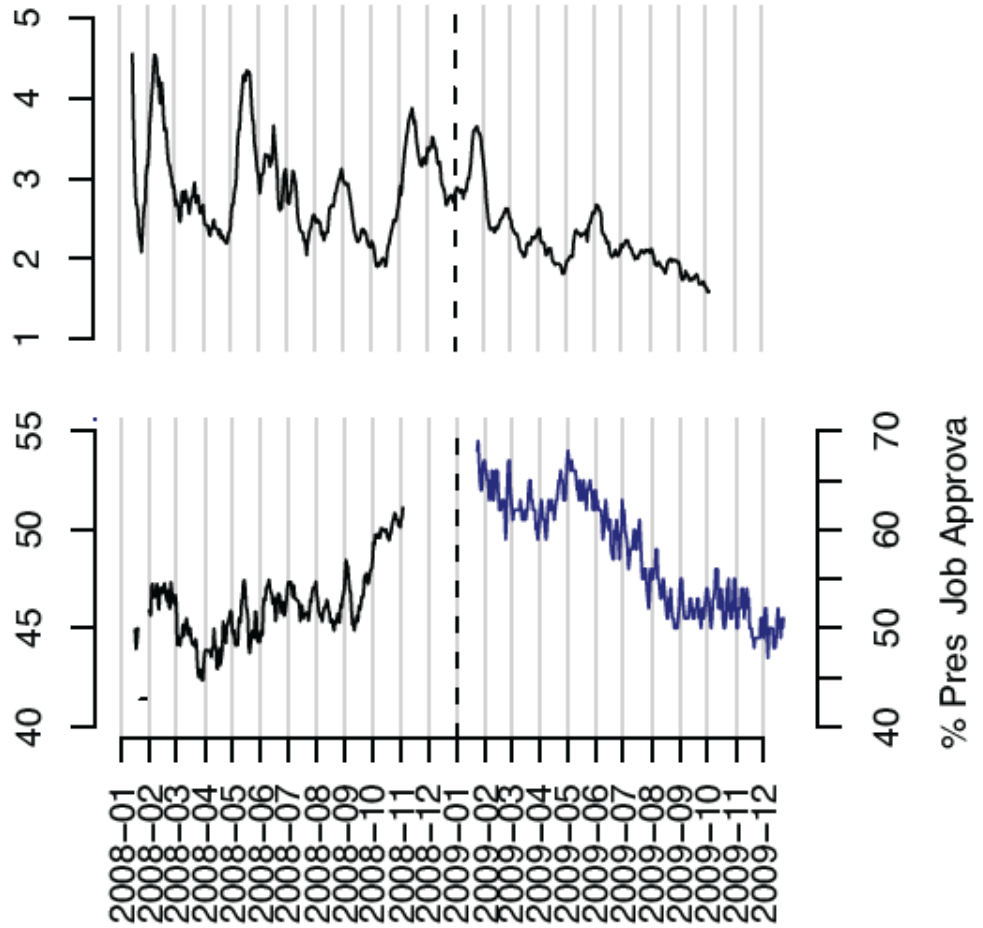

Figure 9: The sentiment ratio for *obama* (15-day window), and fraction of all Twitter messages containing *obama* (day-by-day, no smoothing), compared to election polls (2008) and job approval polls (2009).

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Vì sao cần phân tích ý kiến:

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. PLoS ONE 6(12): e26752. doi:10.1371/journal.pone.0026752

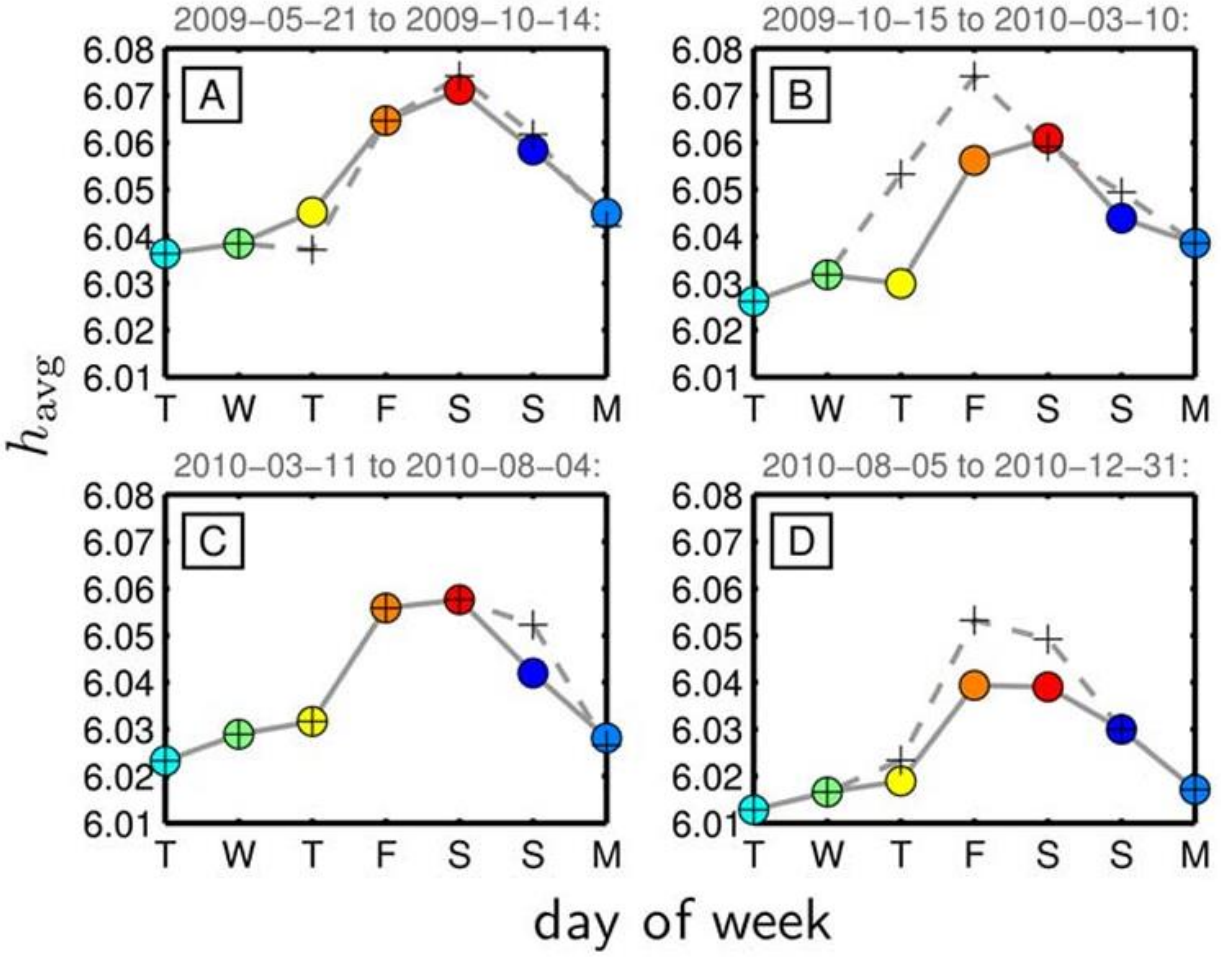

Figure 7. Average of daily average happiness for days of the week over four consecutive time periods of approximately five

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Vì sao cần phân tích ý kiến:

Ý kiến về sở thích, thái độ thể hiện sự đánh giá về:

- Đồ vật: thích Iphone, không thích giày thể thao, v.v…
- Người khác: tín nhiệm chính trị gia, hâm mộ nghệ sĩ, v.v…
- Hành động hay sự kiện: trào lưu thời trang ẩm thực, v.v…

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Các loại yếu tố trong phân tích ý kiến:

1. Chủ thể (holder, source) thể hiện thái độ
2. Đối tượng (target, aspect) của thái độ
3. Loại thái độ
    - Ghét, không thích, trung tính, thích, rất thích.
    - Tiêu cực (negative), trung tính (neutral), tích cực (positive).
4. Đoạn văn bản (text) thể hiện thái độ.

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Các cấp độ phân tích ý kiến:

1. Cấp độ toàn văn bản (document level). Ví dụ: một bài nhận xét (review) về một địa điểm du lịch, một bộ phim.
2. Cấp độ câu (sentence level). Ví dụ: *sản phẩm tốt, đáng 5 sao*.
3. Cấp độ khía cạnh (aspect-based ABSA, feature-level, fine-grained). Ví dụ: ***chụp ảnh sắc nét*** *nhưng* ***mau hết pin***.

1.1 VẤN ĐỀ PHÂN TÍCH Ý KIẾN

Các mức độ phân tích ý kiến:

1. Đơn giản: xác định thái độ trong một đoạn văn bản là tiêu cực hay tích cực.
2. Nâng cao: xác định thái độ theo thang đo nhiều mức độ, chẳng hạn 5 mức độ.
3. Phức tạp: xác định 4 loại yếu tố có trong đoạn văn bản.

1. GIỚI THIỆU

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Bài toán phân tích ý kiến ở mức độ đơn giản và nâng cao:

Cho tập nhãn thái độ $L = \{l_1, l_2, .., l_k\}$ và một đoạn văn bản T, hãy xác định thái độ $l_T \in L$ của đoạn văn bản T.

Ví dụ: L={"*1 Sao*", "*2 Sao*", "*3 Sao*", "*4 Sao*", "*5 Sao*"}. Phân tích ý kiến các bình luận sau:

T_1: *Sản phẩm tạm ổn, nhưng phím bấm khá cứng không có độ nẩy tốt...chất lượng đánh giá sp trung bình*

T_2: *Chất lượng tốt*

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Bài toán phân tích ý kiến ở mức độ đơn giản và nâng cao:

Ví dụ:

★★★★★

2023-08-22 00:26

Sản phẩm tạm ổn, nhưng phím bấm khá cứng không có độ nẩy tốt...chất lượng đánh giá sp trung bình

★★★★★

2023-03-25 13:20

Chất lượng tốt

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Phương pháp giải quyết: Chuyển về bài toán phân lớp:

- Tập nhãn thái độ là tập các phân lớp Y.
- Đoạn văn bản là đầu vào $x \in X$
- Thái độ của đoạn văn bản là phân lớp $y \in Y$ của x.

Xác định hàm phân lớp:

$$f: X \mapsto Y$$

$$f(x) = y$$

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Áp dụng phương pháp học giám sát:

- Tập $Y = \{y_i | i \in [1, k]\}$
- Tập $Train = \{< x_i, y_i > | x_i \in X, y_i \in Y, i \in [1, m]\}$
- Tập $Test = \{< x_i, y_i > | x_i \in X, y_i \in Y, i \in [1, n]\}$
- Sử dụng tập $Train$ để huấn luyện mô hình phân lớp f.
- Sử dụng tập $Test$ để đánh giá hiệu quả của f.
- Các mô hình baseline: Naïve Bayes, Logistic Regression

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Đặc trưng về từ: Các từ điển thái độ

- General Inquirer
- MPQA subjectivity lexicon (Wilson et al. 2005) http://mpqa.cs.pitt.edu/lexicons/
- LIWC (Linguistic Inquiry andWord Count, Pennebaker 2015)
- AFINN (Nielsen 2011)
- NRC Word-Emotion Association
- Lexicon (EmoLex), Mohammadand Turney 2013

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Đặc trưng về từ: Các từ điển thái độ

- General Inquirer (https://inquirer.sites.fas.harvard.edu/homecat.htm)

Positive

ABIDE

H4 Positiv Affil Active Doctrin IAV SUPV

ABILITY

H4Lvd Positiv Strong Virtue EVAL Abs@ ABS MeansLw Noun

ABLE

H4Lvd Positiv Pstv Strong Virtue EVAL MeansLw Modif adjective:

ABOUND

H4 Positiv Passive Increas IAV SUPV

ABSOLVE

H4 Positiv Active SocRel ComForm IAV SUPV

Negative

ABANDON

H4Lvd Negativ Ngtv Weak Fail IAV AffLoss AffTot SUPV

ABANDONMENT

H4 Negativ Weak Fail Noun

ABATE

H4Lvd Negativ Passive Decreas IAV TranLw SUPV

ABDICATE

H4 Negativ Weak Submit Passive Finish IAV SUPV

ABHOR

H4 Negativ Hostile Passive Arousal SV SUPV

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Đặc trưng về từ: Các từ điển thái độ

- MPQA subjectivity lexicon

Positive	**Negative**
Abidance (sự kiên trì)	Abandoned (bỏ rơi)
Abide (tuân thủ)	Abandonment (bỏ rơi)
Abilities (khả năng)	Abandon (bỏ rơi)
Ability (khả năng)	Abase (hạ thấp)
Able (có khả năng)	Abasement (hạ thấp)
…	…

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Đặc trưng về từ: Các từ điển thái độ

- LIWC (Linguistic Inquiry and Word Count, Pennebaker 2015)

II. PSYCHOLOGICAL PROCESSES

Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad

Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Đặc trưng về từ: Các từ điển thái độ

Ví dụ:

1. Thousands of coup supporters **celebrated** overnight, waving flags, blowing whistles …
2. The criteria set by Rice are the following: the three countries in question are **repressive** and **grave human rights violators** …

Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis (Wilson et al., HLT-EMNLP 2005)

1.2 BÀI TOÁN PHÂN TÍCH Ý KIẾN

Các thách thức:

1. Các trường hợp hàm ý, ẩn ý.
2. Cần phải phân tích ngữ cảnh. Ví dụ:

If you **like** it that much, just go ahead and buy it, **like** I did.

MÔ HÌNH NAÏVE BAYES

MỘT SỐ CÔNG THỨC XÁC SUẤT

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

Xác suất biến ngẫu nhiên X có giá trị x

Cho biến ngẫu nhiên $X = \{x_i | i \in [1, k]\}$, xác suất biến ngẫu nhiên X có giá trị x_i, ký hiệu $P(X = x_i)$, thỏa

1. $0 \leq P(X = x_i) \leq 1.$
2. $\sum_{i=1}^{k} P(X = x_i) = 1$

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

Trường hợp xúc xắc cân đối

$$X = \{1,2,3,4,5,6\}$$

$$P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = P(X = 5) = P(X = 6) = \frac{1}{6}$$

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

Trường hợp xúc xắc không cân đối

$$X = \{1,2,3,4,5,6\}$$

Giả sử:

$$P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = P(X = 5) = \frac{4}{25}$$

$$P(X = 6) = \frac{1}{5}$$

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

Bài toán xác định quy luật phân phối

Giả sử quan sát được kết quả tung hai xúc xắc cân đối và không cân đối lần lượt là hai dãy:

1. 1 5 5 5 2 1 6 4 3 3 2 1 6 4 4 2 3
2. 1 2 6 3 5 6 4 1 6 6 3 2 6 5 6 4 6

Hỏi dãy nào là kết quả tung xúc xắc không cân đối.

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

Cho hai biến cố A, B trong phép thử có N kết quả đồng khả năng

- Xác xuất xảy ra cả A và B khi có m kết quả thuận lợi cho AB:

$$P(A,B) = P(A \cap B) = m / N$$

- **Xác suất có điều kiện.** Xác suất xảy ra A khi đã xảy ra B:

$$P(A|B) = P(A,B) / P(B)$$

Ví dụ: Xác suất lần thứ hai lấy được viên bi đỏ khi lần thứ nhất đã lấy được viên bi xanh trong hộp có 3 bi đỏ và 2 bi xanh đồng chất:

$$P(đỏ|xanh) = P(đỏ, xanh) / P(xanh) = (6/20) / (2/5) = 0.75$$

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

- **Quy tắc nhân**. Từ công thức xác suất có điều kiện, ta có:

$$P(A,B) = P(A|B)*P(B) = P(B|A) * P(A)$$

- **Quy tắc phân chia**. Gọi $\bar{B}$ là phần bù của biến cố B, ta có:

$$P(A) = P(A,B) + P(A,\bar{B}) = P(A|B)*P(B) + P(A|\bar{B})*P(\bar{B})$$

- **Công thức Bayes**.

$$P(A|B) = P(B|A)*P(A) / P(B)$$

Trong đó: - P(A|B) được gọi là xác suất hậu nghiệm (Posterior).

- P(A) được gọi là xác suất tiên nghiệm (Prior).

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

- Nếu A là chuỗi $a_1, a_2, \ldots, a_m$, B có các giá trị là $\{b_1, b_2, \ldots, b_n\}$, ta có:

➢ $P(A) = P(a_1, a_2, \ldots, a_m)$

$= P(a_1) \times P(a_2|a_1) \times P(a_3|a_1, a_2) \times \cdots \times P(a_m|a_1, a_2, \ldots, a_{m-1})$

$= P(A, B = b_1) + P(A, B = b_2) + P(A, B = b_3) + \cdots + P(A, B = b_n)$

$= \sum_{i=1}^{n} P(A|\, B = b_i) \times P(B = b_i)$

➢ $P(B|A) = \frac{P(A|B) \times P(B)}{\sum_{i=1}^{n} P(A|B=b_i) \times P(B=b_i)}$

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

- Odds của biến cố B

$$O(B) = \frac{P(B)}{P(\bar{B})} = \frac{P(B)}{1 - P(B)}$$

Nếu P(A) ≥ P(B) thì O(A) ≥ O(B)

2.1 MỘT SỐ CÔNG THỨC XÁC SUẤT

- Điều kiện độc lập

Nếu A và B là hai biến cố độc lập thì:

$$P(A,B) = P(A) \times P(B)$$

$$P(A) = P(A|B)$$

$$P(B) = P(B|A)$$

Tổng quát,

$$P(A_1, A_2, \dots, A_m) = \prod_{i=1}^{m} P(A_i)$$

MÔ HÌNH NAÏVE BAYES

PHÂN TÍCH Ý KIẾN VỚI NAÏVE BAYES

2.2 PHÂN TÍCH Ý KIẾN VỚI NAÏVE BAYES

I like it, like they do

x_1	do	1
x_2	hate	0
x_3	like	1
x_4	love	0
x_5	i	1
x_6	it	1
x_7	they	1

x_1	do	1
x_2	like	2
x_3	i	1
x_4	it	1
x_5	they	1

2.2 PHÂN TÍCH Ý KIẾN VỚI NAÏVE BAYES

Phương pháp phân lớp Naïve Bayes dựa trên công thức Bayes:

$$p(l|d) = \frac{p(l) \times p(d|l)}{p(d)}$$

Khi đó, hàm phân lớp là

$$f(d) = Argmax_{l \in L} p(l|d)$$

$$= Argmax_{l \in L} \frac{p(l)p(d|l)}{p(d)}$$

$$= Argmax_{l \in L} \log\big(p(l)\big) + \log\big(p(d|l)\big)$$

2.2 PHÂN TÍCH Ý KIẾN VỚI NAÏVE BAYES

V={v_1, v_2, .., v_n} là bộ từ vựng, d=$w_1w_2..w_k$, $w_i \in V$, để tính toán

$$p(d|l) = \prod_i p(w_i|l)$$

Các giả thiết độc lập:

- Các từ w_i độc lập với nhau trong tài liệu d.
- Các từ w_i độc lập với vị trí trong tài liệu d.

→ Có hai mô hình Bernoulli Naïve Bayes (Bernoulli NB) và Multinomial Naïve Bayes (Multinomial NB)

2.2 PHÂN TÍCH Ý KIẾN VỚI NAÏVE BAYES

Ưu điểm:

- Tính toán nhanh, dễ dàng.
- Hiệu quả phân lớp văn bản tốt, đặc biệt với tập TRAIN rất lớn.

Nhược điểm:

- Dùng các giả thiết độc lập
- Hiệu quả không cao khi tập TRAIN nhỏ.

MÔ HÌNH NAÏVE BAYES

MULTI-VARIATE BERNOULLI NAÏVE BAYES

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

Mô hình Bernoulli NB: N_l: số tài liệu có nhãn l, N_{l,v_i}: số tài liệu có nhãn l chứa từ $v_i \in V$

- $d=(e_1, e_2, .., e_n)$, với $e_i=1$ nếu $v_i \in d$, $e_i=0$ nếu $v_i \notin d$.
- Ước lượng xác suất prior

$$p(l) = \frac{N_l}{|TRAIN|}$$

- Ước lượng xác suất của từ v_i theo từng lớp l

$$p(v_i|l) = \frac{N_{l,v_i} + 1}{N_l + 2}$$

Smoothing

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

- Tính giá trị

$$p(d|l) = \prod_i p(e_i|l)$$

$$= \prod_{e_i=1} p(v_i|l) \times \prod_{e_i=0} \big(1 - p(v_i|l)\big)$$

- Chọn lớp c thỏa

$$Argmax_{l \in L} \log\big(p(l)\big) + \log\big(p(d|l)\big)$$

hay

$$Argmax_{l \in L} \log\big(p(l)\big) + \sum_{e_i=1} \log(p(v_i|l)) + \sum_{e_i=0} \log(1 - p(v_i|l))$$

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

Ví dụ: Cho tập TRAIN như bên dưới. Cho biết bình luận “i hate it, like they do” là tích cực hay tiêu cực.

Tích cực	Tiêu cực
they love it	they hate it
they like it	hate it, hate it, like it
i love it. they love it	they do hate it
they do love it	

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

V	+ (0.57)	
	1	**0**
do	1	3
hate	0	4
like	1	3
love	3	1
i	1	3
it	4	0
they	4	0

V	- (0.43)	
	1	**0**
do	1	2
hate	3	0
like	1	2
love	0	3
i	0	3
it	3	0
they	2	1

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

V	+ (0.57)	
	1	0
do	1+1	3+1
hate	0+1	4+1
like	1+1	3+1
love	3+1	1+1
i	1+1	3+1
it	4+1	0+1
they	4+1	0+1

V	- (0.43)	
	1	0
do	1+1	2+1
hate	3+1	0+1
like	1+1	2+1
love	0+1	3+1
i	0+1	3+1
it	3+1	0+1
they	2+1	1+1

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

V	+ (0.57)	
	1	0
do	0.33	0.67
hate	0.17	0.83
like	0.33	0.67
love	0.67	0.33
i	0.33	0.67
it	0.83	0.17
they	0.83	0.17

V	- (0.43)	
	1	0
do	0.40	0.60
hate	0.80	0.20
like	0.40	0.60
love	0.20	0.80
i	0.20	0.80
it	0.80	0.20
they	0.60	0.40

2.3 MULTIVARIATE BERNOULLI NAÏVE BAYES

p(+|”i hate it like they do”)

= 0.57 * (0.33 * 0.17 * 0.33 * 0.33 * 0.33 * 0.83 * 0.83)

= 0.000791655

p(-|”i hate it like they do”)

= 0.43 * (0.4 * 0.8 * 0.4 * 0.8 * 0.2 * 0.8 * 0.6)

= 0.004227072

Vậy, bình luận là tiêu cực.

MÔ HÌNH NAÏVE BAYES

MULTINOMIAL NAÏVE BAYES

2.4 MULTINOMIAL NAÏVE BAYES

Mô hình Multinomial NB: N_l: số tài liệu có nhãn l, $\text{count}(d, v_i)$: số lượng từ v_i trong tài liệu d, $v_i \in V$

- Tài liệu $d=w_1w_2..w_k$
- Ước lượng xác suất prior

$$p(l) = \frac{N_l}{|TRAIN|}$$

- Ước lượng xác suất của từ v_i theo từng lớp l

$$p(v_i|l) = \frac{1 + \sum_{<d,l>} count(d, v_i)}{|V| + \sum_{<d,l>} |d|}$$

2.4 MULTINOMIAL NAÏVE BAYES

- Tính giá trị

$$p(d|l) = \prod_i p(w_i|l)$$

- Chọn lớp c thỏa

$$Argmax_{l \in L} \log(p(l)) + \log(p(d|l))$$

hay

$$Argmax_{l \in L} \log(p(l)) + \sum_i \log(p(w_i|l))$$

2.4 MULTINOMIAL NAÏVE BAYES

Ví dụ: Cho tập TRAIN như bên dưới. Cho biết bình luận “i hate it, like they do” là tích cực hay tiêu cực.

Tích cực	Tiêu cực
they love it	they hate it
they like it	hate it, hate it, like it
i love it. they love it	they do hate it
they do love it	

2.4 MULTINOMIAL NAÏVE BAYES

V	+ (0.57)	
do	1	16
hate	0	
like	1	
love	4	
i	1	
it	5	
they	4	

V	- (0.43)	
do	1	13
hate	4	
like	1	
love	0	
i	0	
it	5	
they	2	

2.4 MULTINOMIAL NAÏVE BAYES

V	+ (0.57)	
do	1+1	16+7
hate	0+1	
like	1+1	
love	4+1	
i	1+1	
it	5+1	
they	4+1	

V	- (0.43)	
do	1+1	13+7
hate	4+1	
like	1+1	
love	0+1	
i	0+1	
it	5+1	
they	2+1	

2.4 MULTINOMIAL NAÏVE BAYES

V	+ (0.57)	
do	0.09	1.0
hate	0.04	
like	0.09	
love	0.22	
i	0.09	
it	0.26	
they	0.22	

V	- (0.43)	
do	0.10	1.0
hate	0.25	
like	0.10	
love	0.05	
i	0.05	
it	0.30	
they	0.15	

2.4 MULTINOMIAL NAÏVE BAYES

p(+|”i hate it like they do”)

= 0.57 * (0.09 * 0.04 * 0.26 * 0.09 * 0.22 * 0.09)

= 9.50733E-07

p(-|”i hate it like they do”)

= 0.43 * (0.05 * 0.25 * 0.3 * 0.1 * 0.15 * 0.1)

= 2.41875E-06

Vậy, bình luận là tiêu cực.