



بِهِ نَامِ خَدَا

بخش هشتم

آنالیز احساسات

Sentiment Analysis

حمیدرضا برادران کاشانی



احساسات مثبت و منفی در نظرات فیلم

-  ○ unbelievably disappointing

-  ○ Full of zany characters and richly applied satire, and some great plot twists.

-  ○ This is the greatest screwball comedy ever filmed.

-  ○ It was pathetic. The worst part about it was the boxing scenes.

Hamidreza Baradaran Kashani



Google Product Search



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star 2 3 4 stars 5 stars

What people are saying

ease of use		"This was very easy to setup to four computers."
value		"Appreciate good quality at a fair price."
setup		"Overall pretty easy setup."
customer service		"I DO like honest tech support people."
size		"Pretty Paper weight."
mode		"Photos were fair on the high quality mode."
colors		"Full color prints came out with great quality."

Aspects
(Attributes)

Hamidreza Baradaran Kashani



Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



\$121.53 - \$242.39 (14 stores)

Compare

Average rating (144)

(55)

(54)

(10)

(6)

(23)

(0)

Most mentioned

Performance

Ease of Use

Print Speed

Connectivity

More ▾

Show reviews by source

[Best Buy \(140\)](#)

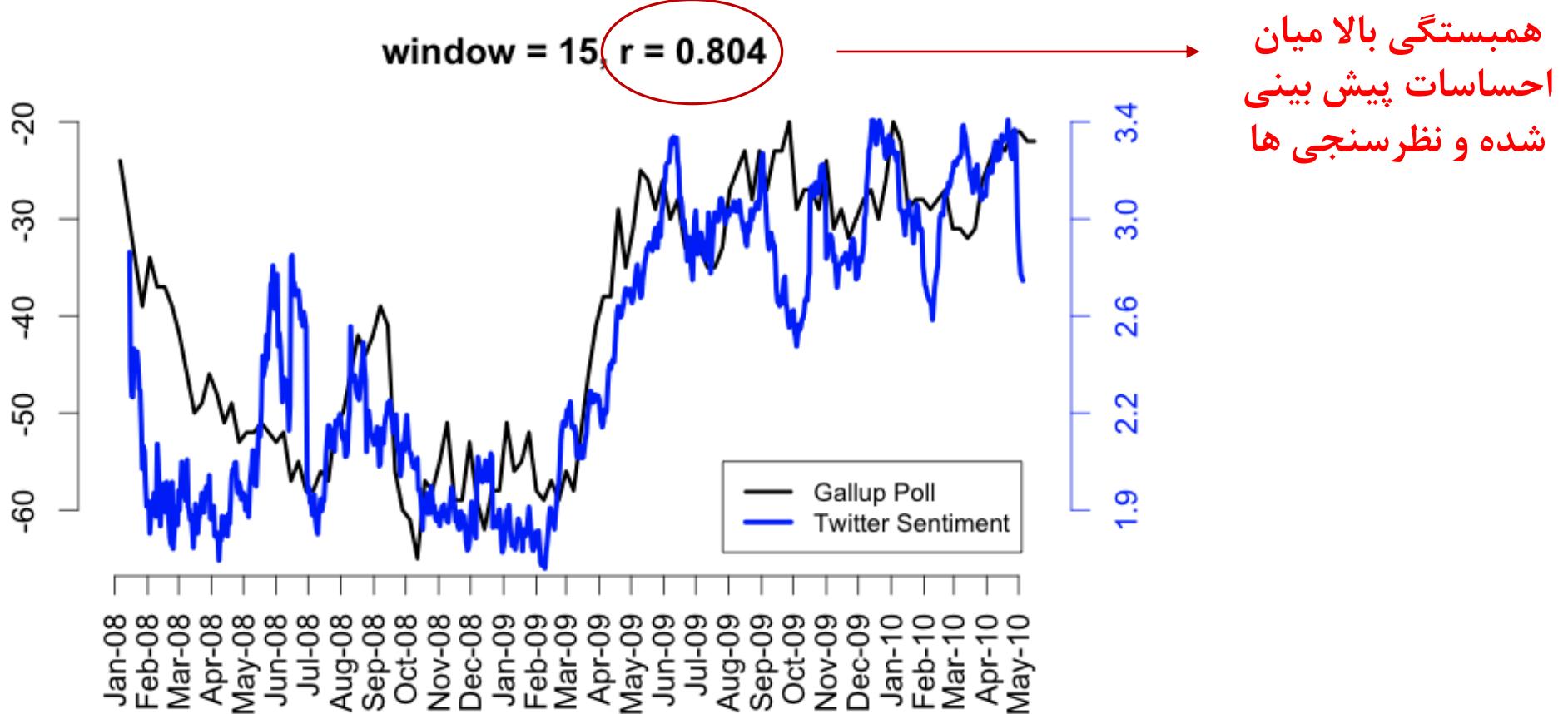
[CNET \(5\)](#)

[Amazon.com \(3\)](#)

radaran Kashani



پیش بینی رضایت مشتریان در توییتر



Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010

Hamidreza Baradaran Kashani



Twitter Sentiment App

Type in a word and we'll highlight the good and the bad

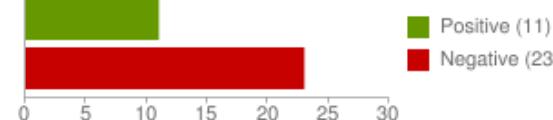
"united airlines" [Save this search](#)

Sentiment analysis for "united airlines"

Sentiment by Percent



Sentiment by Count



jijacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. <http://t.co/Z9QloAjF>
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago



نام های دیگر برای آنالیز احساسات

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis



چرا پردازش احساس؟

- **فیلم**

- آیا نظرات در مورد یک فیلم یا کتاب مثبت است یا خیر؟

- **محصولات**

- نظر مشتریان در مورد گوشی سامسونگ یا کتاب ... چیست؟

- **نظرات عموم مشتریان**

- اعتقاد و رضایت مصرف کنندگان چقدر است؟ آیا نارضایتی در حال افزایش است؟

- **سیاست**

- نظرات مردم جامعه در مورد یک رخداد سیاسی یا مثلاً یک کاندید انتخاباتی به چه سمتی است؟

- **پیش بینی**

- تخمین نتیجه انتخابات یا پیش بینی قیمت سهام و ...

Hamidreza Baradaran Kashani



انواع حالات احساسی / عاطفی

احساس بلندمدت یک شخص : Emotion •

- *angry, sad, joyful, fearful, ashamed, proud, elated*

حال فعلی یک شخص : Mood •

- *cheerful, gloomy, irritable, listless, depressed, buoyant*

موقع عاطفی / احساسی یک شخص نسبت به دیگری در یک تعامل خاص : Interpersonal stance •

- *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*

باور و حس یک شخص در مورد اشیا و افراد اطراف خود : Attitude •

- *liking, loving, hating, valuing, desiring*

شخصیت کلی یک فرد : Personality trait •

- *nervous, anxious, reckless, morose, hostile, jealous*

تعریف آنالیز احساسات



- آنالیز احساسات
- بصورت آشکارسازی نگرش یا نظر شخص یا اشخاص در مورد یک موضوع یا شی یا ...
 - فردی که نظرش را بیان می کند (**Opinion holder**)
 - موضوع یا شی که در مورد آن نظر بیان شود (**Target, aspect**)
 - نوع نظر (**Type**)
 - گروه خاص نظرات (*Like, love, hate, value, desire*)
 - یا نظرات کلی تر و بعضا با شدت آنها (*positive, negative, neutral*)
 - متن شامل نظر
 - یک جمله یا یک سند

Hamidreza Baradaran Kashani



دسته بندی کلی آنالیز احساسات

- مساله ساده
- حس موجود در متن بصورت مثبت یا منفی
- مساله پیچیده تر
- امتیاز دهی به نظر موجود در متن از ۱ تا ۵
- مساله پیشرفتی تر
- فرد گوینده نظر، موضوع یا شی و وجه متناظر با آن و مشخصه های پیچیده تر



یک الگوریتم پایه برای SA

- در نظر گرفتن مساله آنالیز احساس در دامنه فیلم
- حس موجود در متن بصورت مثبت یا منفی
- آشکارسازی قطبیت (**Polarity detection**)
 - آیا نظر نوشته شده در حوزه فیلم (IMDB) مثبت است یا منفی؟
- **polarity Data 2.0** داده •
- <http://www.cs.cornell.edu/people/pabo/movie-review-data>

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

Hamidreza Baradaran Kashani



polarity Data 2.0

- <http://www.cs.cornell.edu/people/pabo/movie-review-data>

Movie Review Data

This page is a distribution site for movie-review data for use in sentiment-analysis experiments. Available are collections of movie-review documents labeled with respect to their overall *sentiment polarity* (positive or negative) or *subjective rating* (e.g., "two and a half stars") and sentences labeled with respect to their *subjectivity status* (subjective or objective) or *polarity*. These data sets were introduced in the following papers:

- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, [Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.](#)
- Bo Pang and Lillian Lee, [A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL 2004.](#)
- Bo Pang and Lillian Lee, [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, Proceedings of ACL 2005.](#)

Until April 2012 (but no longer), we maintained a [list for of other papers using our data](#) the purposes of facilitating comparison of results.

Please cite the version number of the dataset you used in any publications, in order to facilitate comparison of results. Thank you.

Sentiment polarity datasets

- [polarity dataset v2.0](#) (3.0Mb) (includes [README v2.0](#)): 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.
- [Pool of 27886 unprocessed html files](#) (81.1Mb) from which the polarity dataset v2.0 was derived. (This file is identical to movie.zip from data release v1.0.)
- [sentence polarity dataset v1.0](#) (includes [sentence polarity dataset README v1.0](#)): 5331 positive and 5331 negative processed sentences / snippets. Introduced in Pang/Lee ACL 2005. Released July 2005.
- archive:
 - [polarity dataset v1.0](#) (2.8Mb) (includes [README](#)): 700 positive and 700 negative processed reviews. Released July 2002.
 - [polarity dataset v1.1](#) (2.2Mb) (includes [README 1.1](#)): approximately 700 positive and 700 negative processed reviews. Released November 2002. This alternative version was created by [Nathan Treloar](#), who removed a few non-English/incomplete reviews and changing some of the labels (judging some polarities to be different from the original author's rating). The complete list of changes made to v1.1 can be found in [diff.txt](#).
 - [polarity dataset v0.9](#) (2.8Mb) (includes a [README](#)): 700 positive and 700 negative processed reviews. Introduced in Pang/Lee/Vaithyanathan EMNLP 2002. Released July 2002. Please read the "Rating Information - WARNING" section of the README.
 - [movie.zip](#) (81.1Mb): all html files we collected from the IMDb archive.



نمونه ای از پایگاه داده IMDB

when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image. [...]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point.

cool . ★

october sky offers a much simpler image—that of a single white dot , traveling horizontally across the night sky. [...]



“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing. *

it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare.

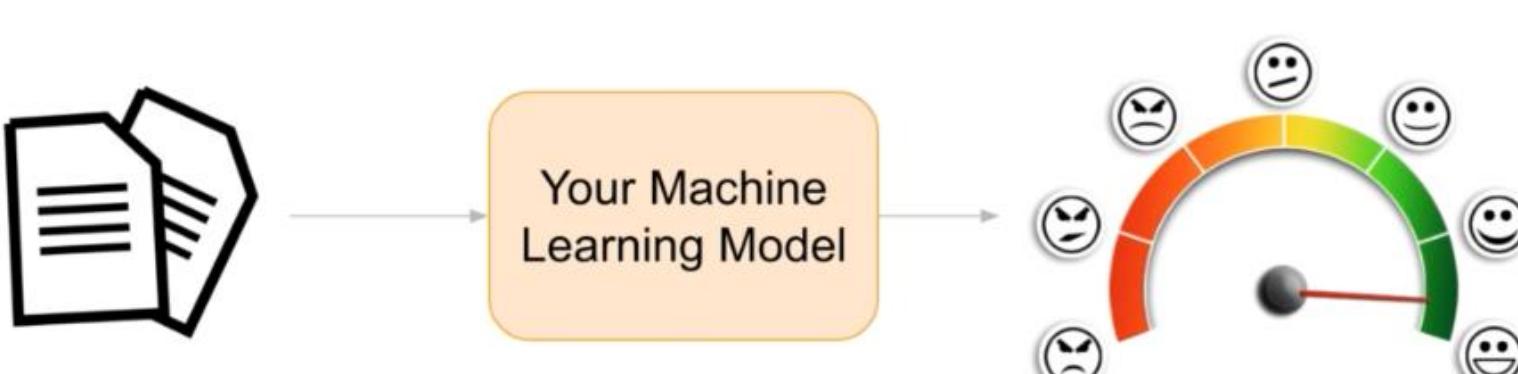
and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents.



Hamidreza Baradaran Kashani



مراحل کلی یک الگوریتم پایه



❖ دسته بندی نظرات با کلاسیفایرهاي مختلف (یادگيري بانظارت)

❖ توکن بندی

❖ استخراج ویژگی

❖ بیز ساده

❖ بیشینه آنتروپی

SVM

Logistic Regression

شبکه های عصبی

...



چالش های موجود در توکن بندی SA

- ❖ با فرمات های HTML و XML چگونه برخورد کنیم؟
 - ❖ وجود برخی نشانه ها در توییت ها (مثلا نام افراد و هشتگ ها و ...)
 - ❖ کلمات با حرف اول بزرگ یا تمام حروف بزرگ
 - ❖ تاریخ ها و شماره های تلفن و ..
 - ❖ وجود Emoticon ها
 - ❖ برخی کدهای مربوط به توکن بندی برای SA
- [Christopher Potts sentiment tokenizer](#)
 - [Brendan O'Connor twitter tokenizer](#)



استخراج ویژگی ها برای SA

❖ نحوه برخورد با نفی (negation) در جمله

- I **didn't** like this movie
VS
- I really like this movie

❖ کدام نوع کلمات در نظر گرفته شوند؟

- ❖ صفت ها یا تمام کلمات؟
- ❖ معمولاً تمام کلمات نتیجه بهتری می دهند.



Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).
Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

❖ راه حل: به هر کلمه بعد از نفی تا اولین علامت نقطه گذاری بعدی، عبارت NOT را اضافه کنید

didn't like this movie , but I



didn't NOT_like NOT_this NOT_movie but I

❖ چالش بیشتر در جمله: "نمی خواهم بگم که این فیلم را دوست ندارم" ???



بیز ساده (Naïve Bayes)

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \tilde{\bigcirc} \prod_{i \in positions} P(w_i | c_j)$$

$$\hat{P}(w | c) = \frac{count(w, c) + 1}{count(c) + |V|}$$



بیز ساده دودویی (Binarized Multinomial Naïve Bayes)

- ❖ برای آنالیز احساسات و برخی کاربردهای دیگر دسته بندی متن:
- ❖ رخداد کلمه در متن مهمتر از تعداد فراوانی آن است.
- مثلاً رخداد کلمه "جذاب" در متن اطلاعات زیادی در ارتباط با احساس موجود در متن می‌دهد،
- اما اینکه کلمه "جذاب" سه یا ۴ بار در متن ظاهر شده است اطلاعات چندانی ندارد!
- ❖ روش بیز ساده دودویی یا BMNB:
- ❖ این روش به جای تعداد تکرار کلمات از ۰ یا ۱ به معنای عدم حضور یا حضور کلمه استفاده می‌کند.



روش BMNB: مرحله آموزش

- ❖ مرحله ۱) استخراج واژگان ۷ از پیکره یادگیری (از کل C کلاس)
❖ اندازه واژگان = تعداد لغات = $|V|$

❖ مرحله ۲) محاسبه احتمالات پیشین کلاس ها $P(c_j)$

❖ تقسیم تعداد سندهای متعلق به هر کلاس به تعداد کل سندها

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

$docs_j \leftarrow \text{all docs with class } = c_j$

❖ مرحله ۳) حذف تکرارها از هر سند

❖ نگه داشتن تنها یک نمونه از هر کلمه

Hamidreza Baradaran Kashani



روش BMNB: مرحله آموزش

- ❖ مرحله ۳) محاسبه احتمالات شباهت $P(w_k | c_j)$
- ❖ تشکیل یک فایل متنی حاوی کلیه متون آموزش متعلق به کلاس C_j

$Text_j \leftarrow$ single doc containing all $docs_j$

برای هر کلمه در واژگان V

تعداد رخدادهای کلمه w_i در متن $Text_j$

$n_j \leftarrow$ # of all words in $Text_j$

$n_{ij} \leftarrow$ # of occurrences of w_i in $Text_j$

$\hat{p}(w_i | c_j) = \frac{n_{ij} + \alpha}{n_j + \alpha |V|}$ محاسبه احتمال $P(w_i | c_j)$

Hamidreza Baradaran Kashani



روش BMNB: مرحله آزمون

- ❖ حذف تمام کلمات تکراری از سند d
- ❖ استفاده از رابطه بیز ساده برای تعیین کلاس نظر (مثبت، منفی، ...)

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \tilde{\bigcirc} \prod_{i \in positions} P(w_i | c_j)$$



مثال (مقایسه بیز ساده نرمال و دودیی)

Normal	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Boolean	Doc	Words	Class
Training	1	Chinese Beijing	c
	2	Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Tokyo Japan	?

روش بیز ساده دودویی



-
- B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.
- V. Metsis, I. Androutsopoulos, G. Palioras. 2006. Spam Filtering with Naive Bayes – Which Naive Bayes? CEAS 2006 - Third Conference on Email and Anti-Spam.
- K.-M. Schneider. 2004. On word frequency information and negative evidence in Naive Bayes text classification. ICANLP, 474-485.
- JD Rennie, L Shih, J Teevan. 2003. Tackling the poor assumptions of naive bayes text classifiers. ICML 2003

❖ بر اساس مقالات فوق، روش باینری کارایی بهتری نسبت به تکرار کلمات دارد.

❖ روش دیگر (Rennis 2003): استفاده از لگاریتم تعداد تکرار یک کلمه ($\log(freq(w))$)
○ می تواند عددی کمتر از فراوانی اما متفاوت با مقدار ۱ باشد.

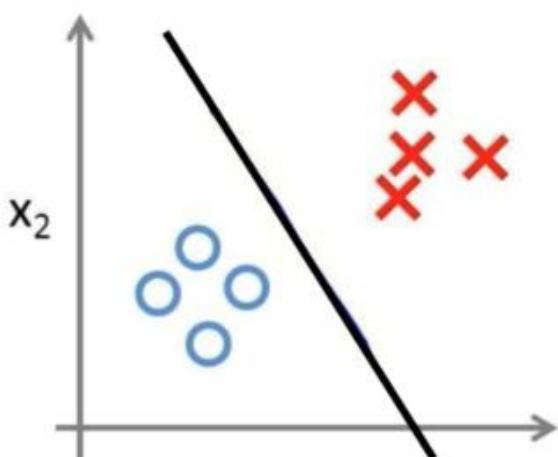
❖ روش های SVM و MaxEnt بهتر از بیز ساده عمل می کنند.



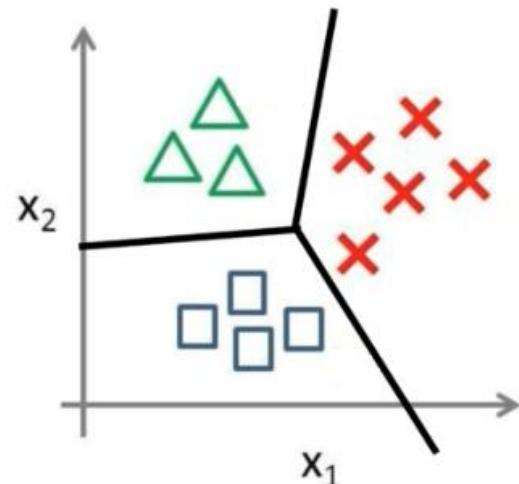
روش Logistic Regression

- ❖ روشن LR یک روش دسته بندی خطی است.
- ❖ برخلاف روش بیز ساده که یک دیدگاه احتمالاتی برای حل مساله دسته بندی دارد، روش LR مبتنی بر عملیات در فضای برداری است.
- ❖ تبدیل هر نمونه به یک بردار ویژگی و بدست آوردن یک **مرز خطی** بصورت خط (یا صفحه با ابرصفحه) که کلاسهها را از یکدیگر جدا کند.

Binary classification:



Multi-class classification:

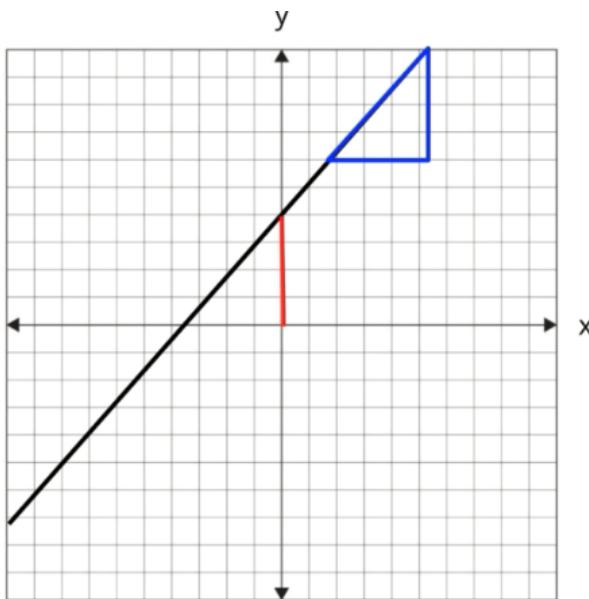


Hamidreza Baradaran Kashani



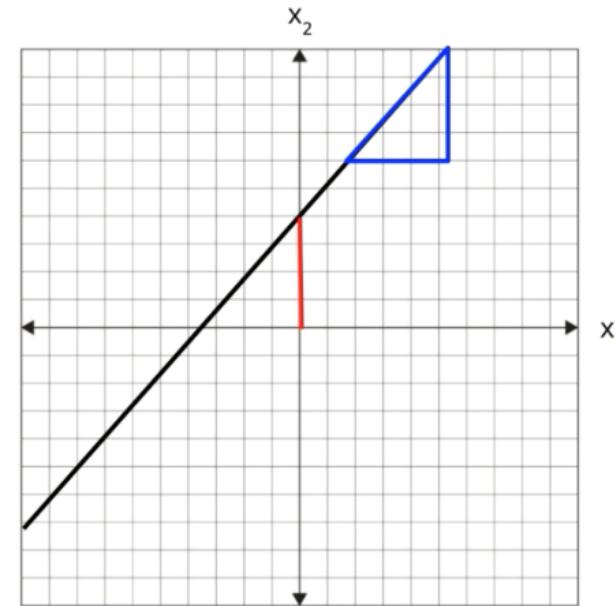
روش Logistic Regression

❖ معادله یک خط (به عنوان مرز تصمیم یا همان دسته بند)



$$y = \boxed{m}x + \boxed{b}$$

Slope Intercept



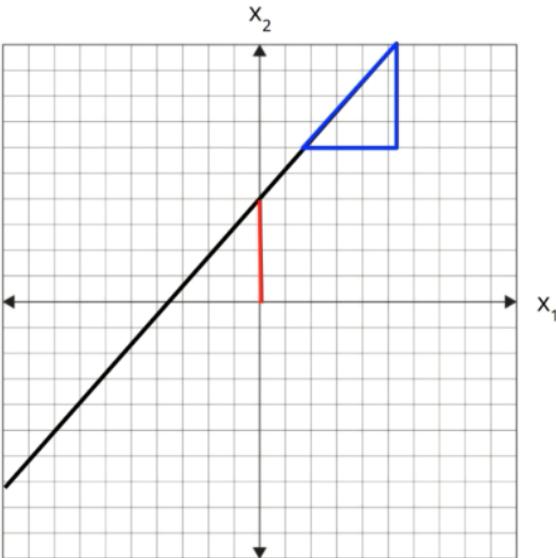
$$x_2 = \boxed{m}x_1 + \boxed{b}$$

Slope Intercept

Hamidreza Baradaran Kashani



روش Logistic Regression



❖ معادله یک خط (به عنوان مرز تصمیم یا همان دسته بند)

$$w_1x_1 + w_2x_2 + b = 0$$

Weights Bias

$$w_1x_1 + w_2x_2 + b = 0$$

$$x_2 = \left(\frac{-w_1}{w_2} \right) x_1 + \left(\frac{-b}{w_2} \right)$$

Slope Intercept

❖ LR مثل یک نرون (neuron) عمل می کند (نرون واحد پایه در یک شبکه عصبی است)

Hamidreza Baradaran Kashani



روش Logistic Regression

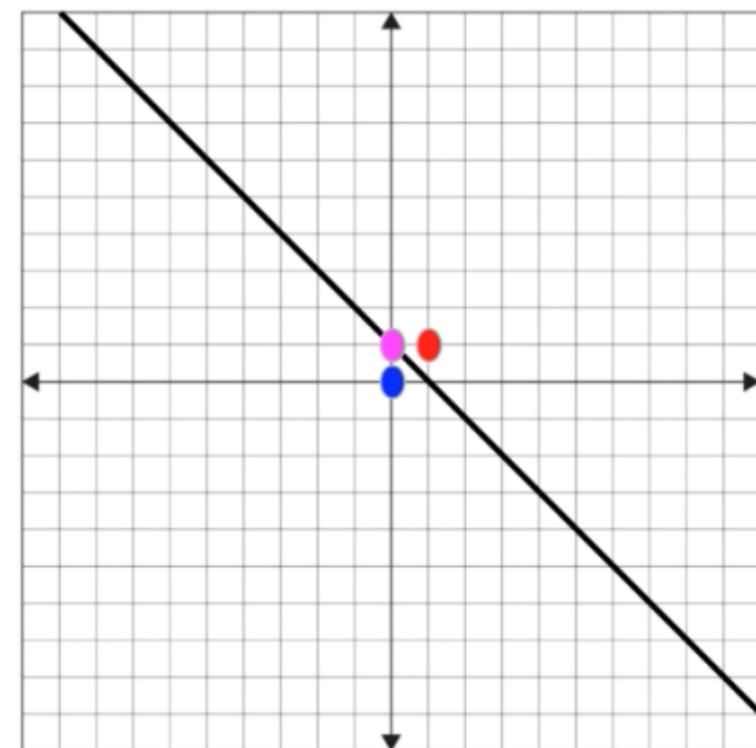
❖ معادله یک خط (به عنوان مرز تصمیم یا همان دسته بند)

$$x_1 + x_2 - 1 = 0$$

$x=(0,1) \rightarrow \text{on the line}$

$x=(0,0) \rightarrow \text{left of the line}$

$x=(1,1) \rightarrow \text{right of the line}$



Hamidreza Baradaran Kashani



روش Logistic Regression

❖ نمایش برداری معادله مرز تصمیم

- در عمل همواره بعد بردارهای ویژگی استخراجی از متون بیشتر از ۲ است مثلا ۳۰۰ بعد ($D=300$)
- استفاده از یک نمایش برداری برای مرز تصمیم در این موارد

$$x = (x_1, x_2, \dots, x_D)^T, w = (w_1, w_2, \dots, w_D)^T$$

$$a(x) = \sum_{i=1}^D w_i x_i + b = w^T x + b$$

$a(x) = 0$: on the line

$a(x) < 0$: one side of the line

$a(x) > 0$: other side of the line

Activation



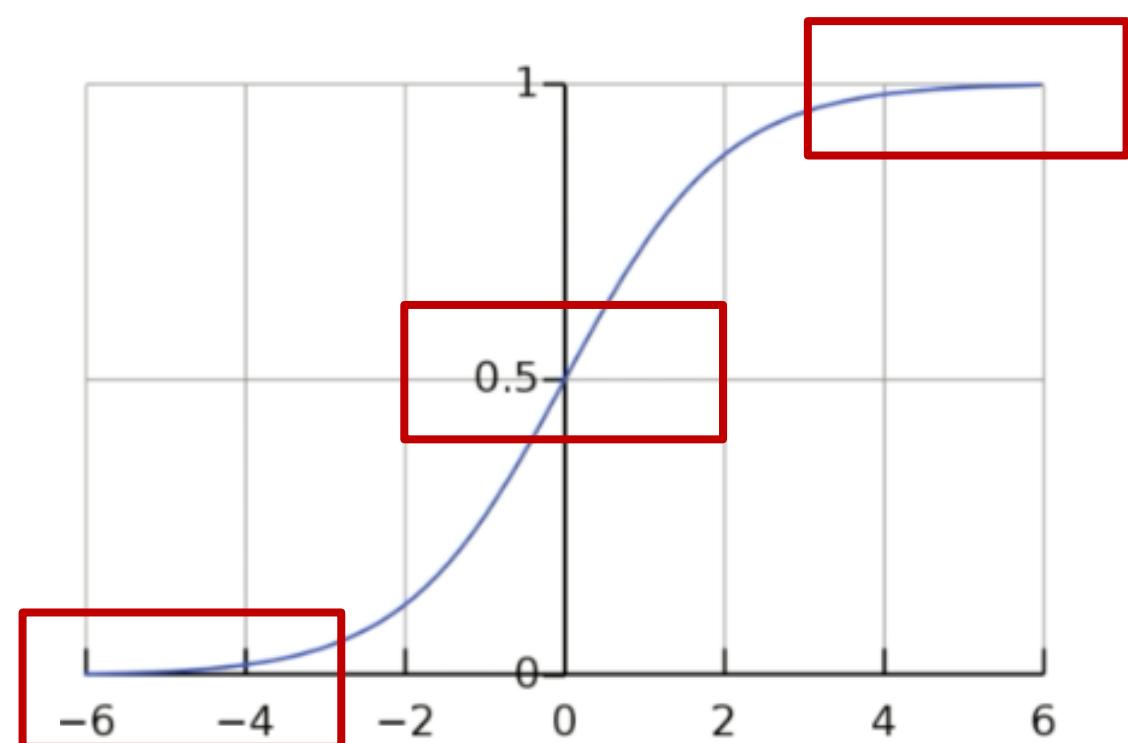


روش Logistic Regression

تابع Sigmoid یا Logistic ❖

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

خواص تابع سیگموید؟

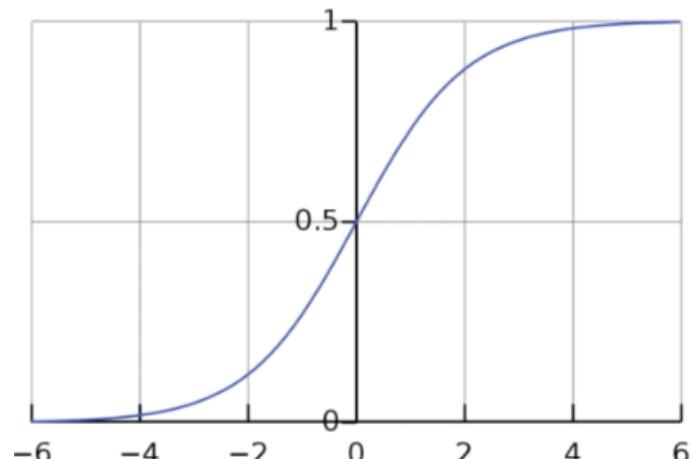




روش Logistic Regression

❖ تابع Sigmoid یا Logistic

❖ تبدیل خروجی کلاسیفایر به یک مقدار احتمالاتی (عددی بین ۰ و ۱) خروجی تابع سیگموید عددی بین ۰ و ۱ است.



$$p(y = 1 | x) = \sigma(w^T x + b)$$

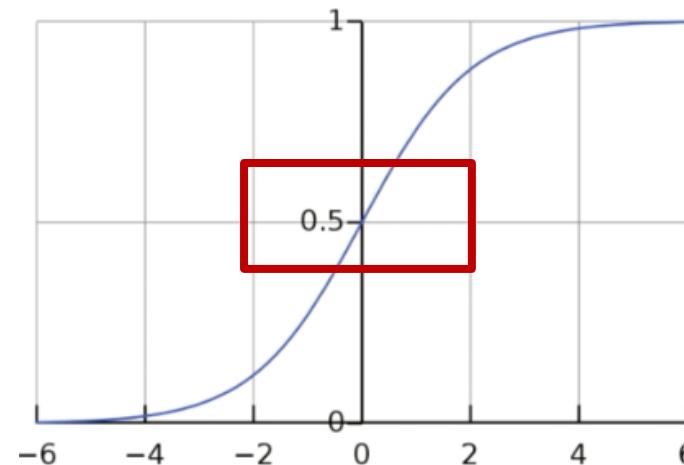
$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$



روش Logistic Regression

❖ قاعده تصمیم‌گیری در LR

Prediction =
round (p (y=1|x))



	$a(x) = w^T x + b$	$p(y=1 x)$
Predict 1	$a(x) > 0$	$p(y=1 x) > 0.5$
Predict 0	$a(x) < 0$	$p(y=1 x) < 0.5$
Predict ? (Usually 1)	$a(x) = 0$	$p(y=1 x) = 0.5$

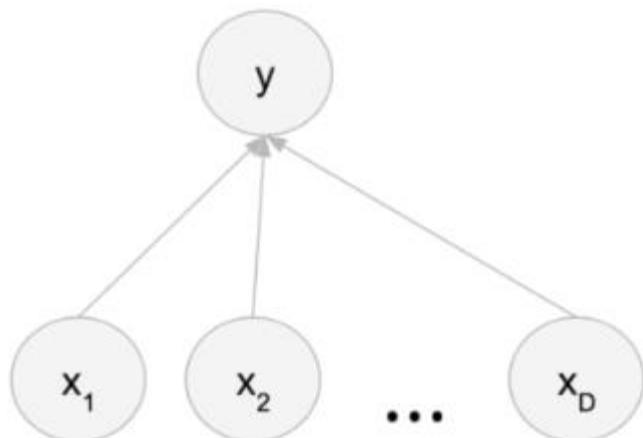
Hamidreza Baradaran Kashani



Naïve Bayes با Logistic Regression مقایسه

Logistic Regression

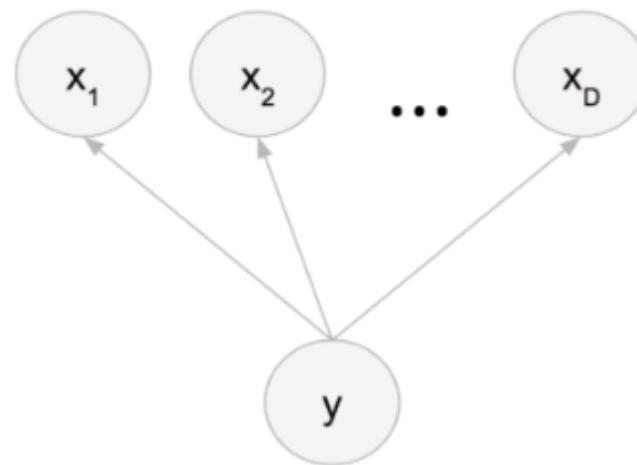
$$p(y | x)$$



Discriminative model

Naïve Bayes

$$p(x | y)$$



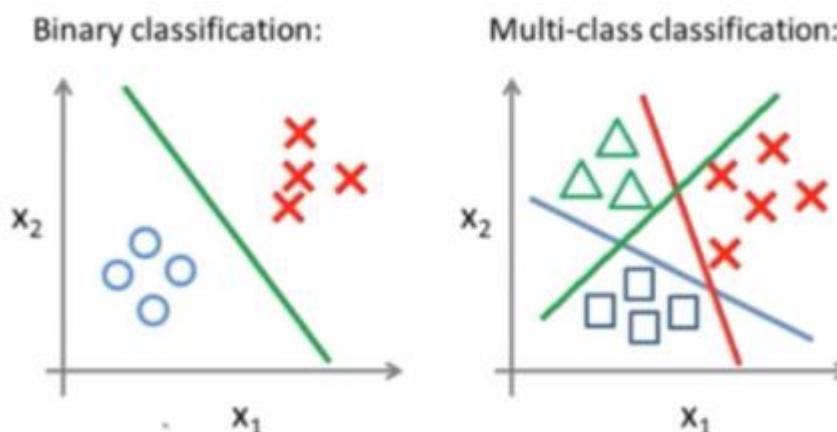
Generative model

Hamidreza Baradaran Kashani



روش Logistic Regression چند کلاسی

- ❖ نام های دیگر (خیلی خوب نیستند!!):
 - Multinomial Logistic regression
 - Maximum Entropy Classifier
- ❖ همان دسته بند LR است که برای مواردی که بیشتر از دو کلاس داریم استفاده می شود.



Hamidreza Baradaran Kashani



روش Logistic Regression چند کلاسی

$$w_1, w_2, \dots, w_K$$

$$a_1 = w_1^T x + b_1$$

$$a_2 = w_2^T x + b_2$$

...

$$a_K = w_K^T x + b_K$$

❖ فرض کنیم K کلاس داریم:

- در اینصورت K تا بردار w داریم که هر کدام یک بردار ستونی D^*1 بعدی هستند:

b_1, b_2, \dots, b_K همین طور K مقدار بایاس داریم:

- حال می توان برای یک بردار ویژگی ورودی x که یک بردار ستونی D^*1 بعدی است، K مقدار خروجی یا K مقدار a_1, a_2, \dots, a_K را محاسبه کرد:



روش Logistic Regression چند کلاسی

تابع Softmax

❖ در ادامه به جای تابع سیگموید که برای حالت دو کلاسی استفاده می شود، از تابع Softmax استفاده می شود.

$$\text{softmax}(a)_k = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)}, \text{ for } k = 1 \dots K$$

❖ مشابه با تعبیری که برای خروجی تابع سیگموید داشتیم، خروجی Softmax را هم می توان بصورت احتمال تعریف کرد:

$$p(y = k | x) = \text{softmax}(a)_k$$

❖ جمع احتمال تعلق یک نمونه به تمام K کلاس برابر با یک است.

❖ در واقع تابع Softmax مقادیر activations را تبدیل به یک توزیع احتمالاتی بر روی K کلاس می کند.



روش Logistic Regression چند کلاسی

❖ تابع Softmax

- ❖ پس برای هر نمونه، K مقدار خروجی از تابع Softmax داریم.
- ❖ اگر N نمونه داشته باشیم، خروجی نهایی یک ماتریس (با مفهوم احتمالات) با بعد $N \times K$ است (هر ردیف بردار بیانگر احتمالات تعلق یک نمونه به K کلاس مختلف است).
- ❖ برای بدست آوردن پیش بینی کلاس نهایی برای هر نمونه:

$$k^* = \arg \max_k P(Y = k | X)$$



تعبیر مدل Logistic Regression باینری

- ❖ یادگیری روش LR یک فرآیند تکرار شونده است.
- ❖ وزن های بدست آمده از فرآیند یادگیری LR قابل تفسیر و تعبیر هستند (یک مشخصه مثبت روش LR).
- ❖ فرض: ویژگی های ورودی در X غیرمنفی هستند.
- ❖ در برخی ویژگی های استخراجی از متن مثل BoW و ... که مبتنی بر فراوانی هستند این مساله صادق هست.
- ❖ برای ورودی های منفی این تفسیرها را معکوس بیان می کنیم.

مثال برای LR باینری



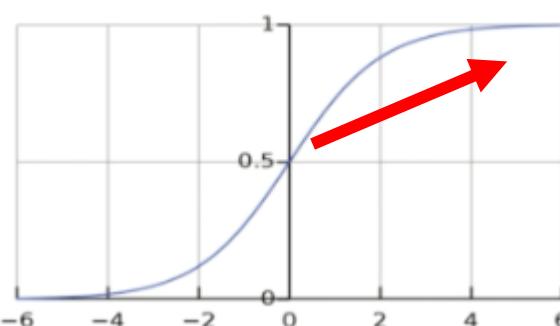
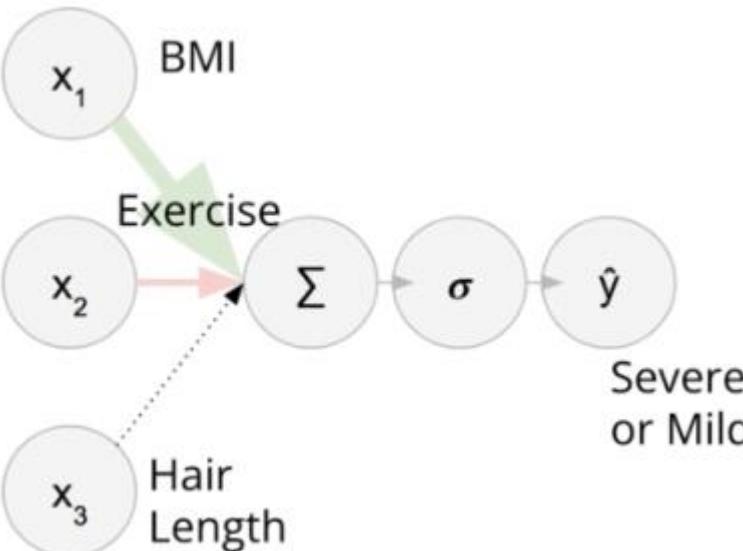
❖ ورودی: داده ها شامل ۳ مولفه هستند: x_1 , x_2 , x_3 , Exercise , BMI , Hair length , Frequency

❖ خروجی: پیش بینی آن که آیا فرد بیماری کووید را تجربه کرده است یا خیر؟ (بیماری شدید یا خفیف)

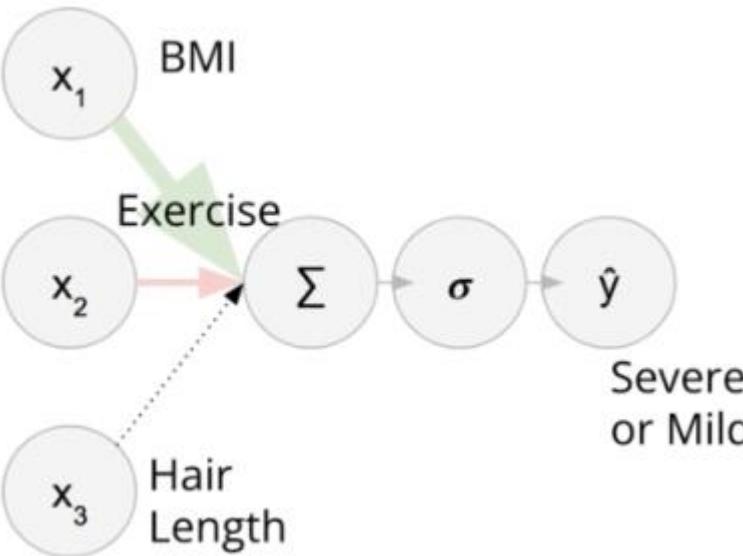
$w_{\text{BMI}} = 2$ (وزن تخصیص داده شده به ویژگی BMI بزرگ و مثبت است)

❖ بزرگ بودن وزن مقدار activation بزرگتری هم ایجاد می کند و این باعث می شود احتمال به سمت ۱ برود.

❖ تعبیر: وزن های بزرگ اثر مثبت زیادی روی ۱ بودن خروجی دارد.



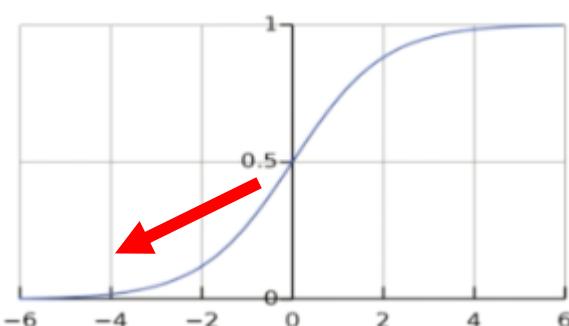
مثال برای LR باینری



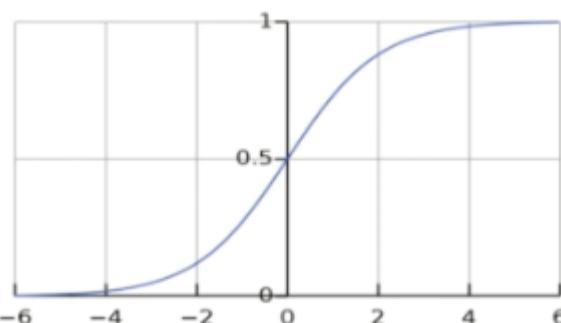
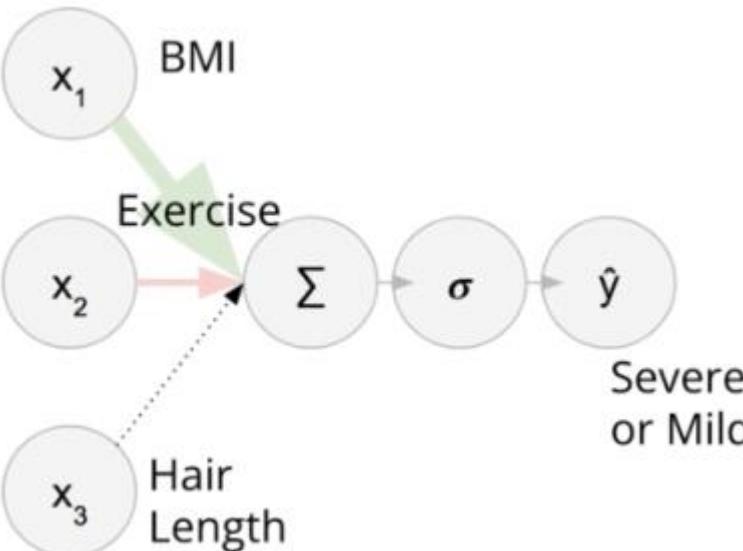
وزن تخصیص داده شده به ویژگی دوم منفی و کوچک است) $W_{Exercise} = -0.5$

ضرب وزن منفی در ویژگی ورودی با مقدار مثبت، مقدار activation را به سمت منفی می برد و این باعث می شود مقدار احتمال به سمت ۰ برود.

البته چون مقدار وزن کوچک است تاثیر کمتری نسبت به یک وزن بزرگ دارد.



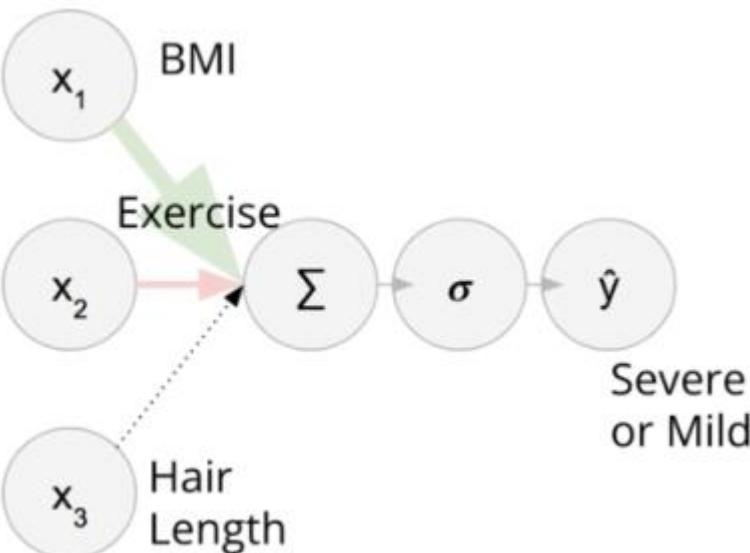
مثال برای LR باینری



- ❖ مثال:
- ❖ دو عامل مهم: **دامنه و علامت وزن ها**
- ❖ **تعییر وزن با علامت مثبت**
 - خروجی ها به سمت ۱ می رود و ویژگی های ورودی همبستگی با کلاس مثبت دارند
- ❖ **تعییر وزن با علامت منفی**
 - خروجی ها به سمت ۰ می رود و ویژگی های ورودی همبستگی با کلاس منفی دارند.
- ❖ **تعییر دامنه یا اندازه وزن**
 - هر چه وزن مقدار مطلق بزرگتر داشته باشد، تاثیرش در تغییر مقدار خروجی بیشتر است.

Hamidreza Baradaran Kashani

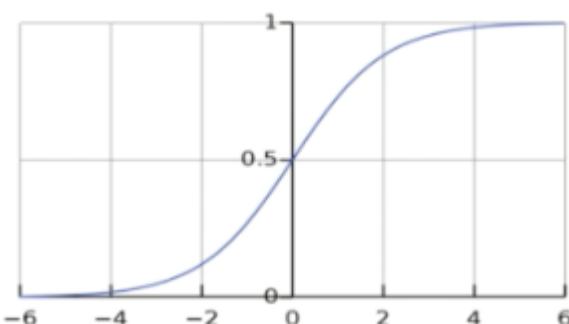
مثال برای LR باینری



(وزن تخصیص داده شده به ویژگی سوم صفر است) $w_{HairLength} = 0$

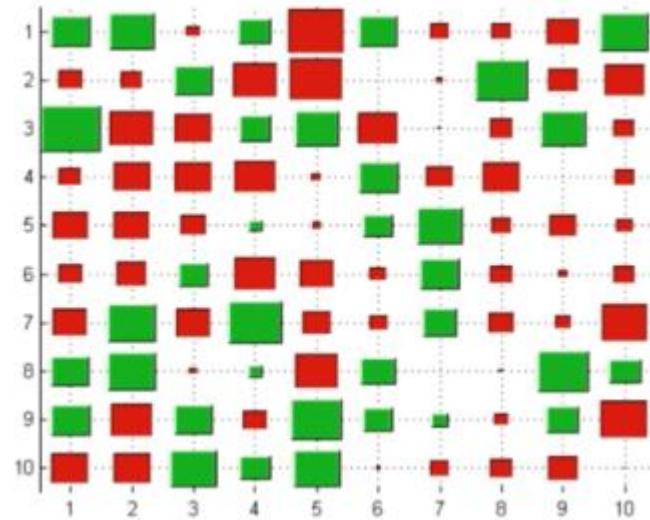
❖ تعبیر وزن صفر:

- ویژگی سوم ویژگی نامربوط (irrelevant) است.
- یعنی اثری در خروجی ندارد.





مثال برای LR چند کلاسی



- ❖ در حالت چند کلاسی مساله پیچیده تر است، چون:
 - چرا که تعداد وزن ها بیشتر است
 - از طریق تابع Softmax به خروجی مرتبط می شوند.

$$W = \begin{bmatrix} w_1^T \\ w_2^T \\ \dots \\ w_K^T \end{bmatrix}_{D^*K}$$

$$w_1, w_2, \dots, w_K$$

- حال یک ماتریس W بصورت زیر ایجاد می کنیم که بعد آن D^*K است.



مثال برای LR چند کلاسی

- ❖ اگر $W_{d,k}$ (مولفه ردیف d و ستون k ماتریس) مقدار **بزرگ** و مثبت باشد:
 - یعنی ویژگی d -ام همبستگی مثبت زیادی با کلاس k -ام دارد.

- ❖ اگر $W_{d,k}$ مقدار **بزرگ** و **منفی** باشد:
 - یعنی ویژگی d -ام همبستگی منفی زیادی با کلاس k -ام دارد.

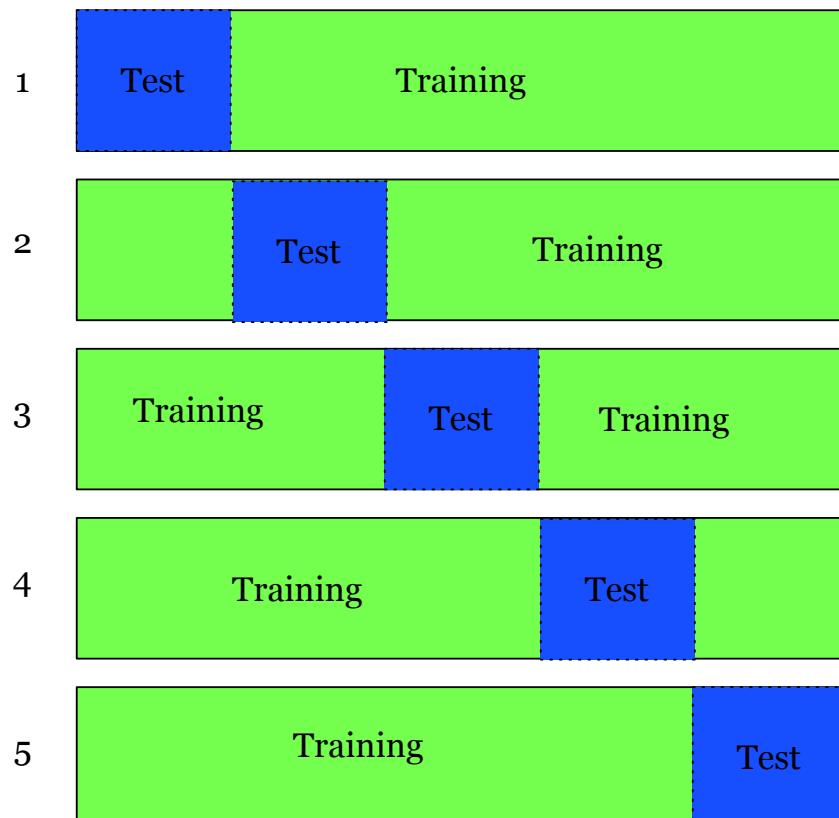
- ❖ دامنه وزن بزرگتر $:W_{d,k}$
 - تاثیر بیشتر ویژگی d برای آن که یک نمونه به کلاس k متعلق باشد (علامت وزن +) یا نباشد (علامت وزن -).

- ❖ دامنه وزن کوچکتر
 - تاثیر کمتر ویژگی d برای آن که یک نمونه به کلاس k متعلق باشد (علامت وزن +) یا نباشد (علامت وزن -).



روش ارزیابی Cross-Validation

Iteration



❖ تقسیم دادگان مثلا به ۵ یا ۱۰ بخش مساوی (ترجیحا نسبت کلاس ها مثبت و منفی در تمام بخش ها یکسان باشد)

❖ برای هر بخش (fold)

- یادگیری یک مدل (یا دسته بند) بر روی ۹ فولد یادگیری تست مدل بر روی تک فولد تست و محاسبه عملکرد مثلا دقیق بر روی آن فولد
- ❖ متوسط گیری عملکرد ۱۰ اجرا بر روی ۱۰ فولد تست



برخی از جملات پیچیده در SA

- ❖ “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”

- ❖ “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.”

- ❖ Well as usual Keanu Reeves is nothing special, but surprisingly, the very talented Laurence Fishbourne is not so good either, I was surprised.

Hamidreza Baradaran Kashani



Sentiment Lexicons

Hamidreza Baradaran Kashani



چرا **Sentiment Lexicon**؟

- ❖ واژگان یا لغت نامه احساس (**Sentiment lexicon**) منابع زبانی هستند که در آن کلمات مختلف که بار احساسی مثبت و منفی دارند مشخص شده اند.
- ❖ بعضاً شدت مثبت یا منفی بودن هر کلمه نیز مشخص شده است.
- ❖ در برخی موارد به آن **Polarity lexicon** هم می گویند.



The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. **The General Inquirer**: A Computer Approach to Content Analysis. MIT Press

یکی از واژگان احساس مشهور در زبان انگلیسی 

- Home page: <http://www.wjh.harvard.edu/~inquirer>
- List of Categories: <http://www.wjh.harvard.edu/~inquirer/homecat.htm>
- Spreadsheet: <http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls>

Categories:

- Positive (1915 words) and Negative (2291 words)
- Strong vs Weak, Active vs Passive, Overstated versus Understated
- Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc

Free for Research Use

Hamidreza Baradaran Kashani



LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

Home page: <http://www.liwc.net/>

2300 words

Affective Processes

- negative emotion (*bad, weird, hate, problem, tough*)
- positive emotion (*love, nice, sweet*)

Cognitive Processes

- Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)

Pronouns, Negation (*no, never*), Quantifiers (*few, many*)

\$30 or \$90 fee

Hamidreza Baradaran Kashani



MPQA Subjectivity Cues Lexicon

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

Home page: http://www.cs.pitt.edu/mpqa/subj_lexicon.html

6885 words from 8221 lemmas

- 2718 positive
- 4912 negative

Each word annotated for intensity (strong, weak)

GNU GPL

Hamidreza Baradaran Kashani



Bing Liu Opinion Lexicon

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.

[Bing Liu's Page on Opinion Mining](#)

<http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>

6786 words

- 2006 positive
- 4783 negative

Hamidreza Baradaran Kashani



SentiWordNet

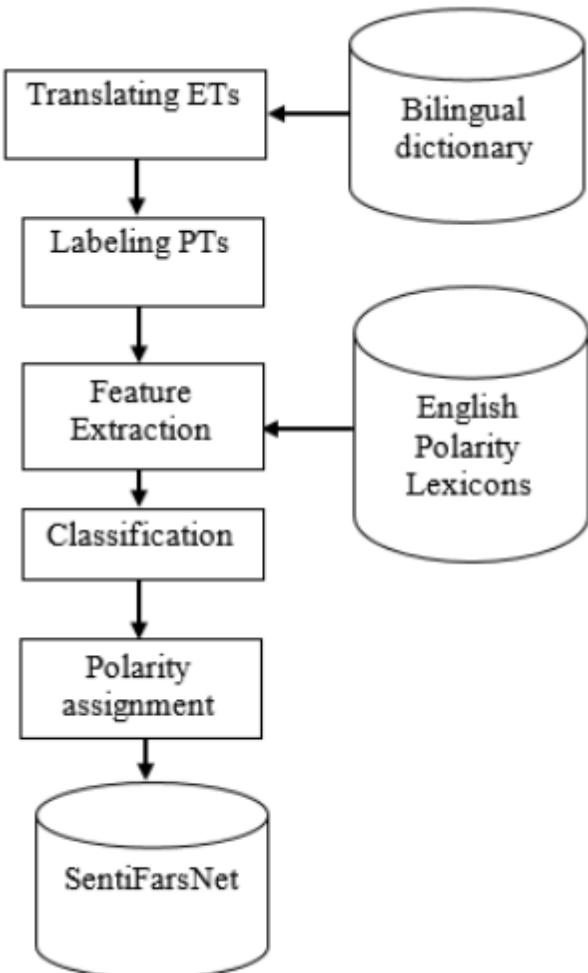
Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

- Home page: <http://sentiwordnet.isti.cnr.it/>
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
- [estimable(J,3)] “may be computed or estimated”
 Pos 0 Neg 0 Obj 1
- [estimable(J,1)] “deserving of respect or high regard”
 Pos .75 Neg 0 Obj .25

Hamidreza Baradaran Kashani



SentiFars



Rahim Dehkharghani. 2019. SentiFars: A Persian Polarity Lexicon for Sentiment Analysis. 1, 1, Article 1 (July 2019), 12 pages.

استفاده از ۴ واژگان قطبیت انگلیسی برای ساخت SentiFars

- NRC Emotion Lexicon
- SentiWordNet 3.0
- SenticNet 3.0
- Liu's Polarity Lexicon



SentiFars

Feature name

- f_1 : polarity score of ET in SenticNet
- f_2 : positive polarity score of ET in SWN
- f_3 : negative polarity score of ET in SWN
- f_4 : positive polarity label of ET in NRC lexicon
- f_5 : negative polarity label of ET in NRC lexicon
- f_6 : polarity label of ET in Liu's lexicon

Persian term	شگفت‌انگیز			
English term	Wonderful			
Polarity Lexicons	SN	SWN(P,N)	NRC(P,N)	Liu
Polarity scores	0.355	(0.75,0)	(1,0)	1
SF scores	$[Pos, Obj, Neg] = [0.875, 0.125, 0]$			

Fig. 4. A sample Persian term in SentiFars, generated by four English polarity lexicons.

Hamidreza Baradaran Kashani



مقایسه با سایر منابع قطبیت موجود در فارسی SentiFars

Polarity lexicon/Corpus	Granularity level	Size(P,O,N)	Scoring form
SentiFars	Word/Phrase	(724,819,1153)	(P,O,N) scores
PerSent	Word/Phrase	(203,986,202)	Overall score
LexiPers	Word/Phrase	(995,4573,1335)	Polarity label
SentiPers	Aspect/Sentence /Document	(21471,1661,3864)	-2 to +2
Lexicon of [9]	Word	(1761, 0, 2150)	Polarity label

Hamidreza Baradaran Kashani



عدم توافق بین ها Polarity Lexicon

Christopher Potts, [Sentiment Tutorial](#), 2011

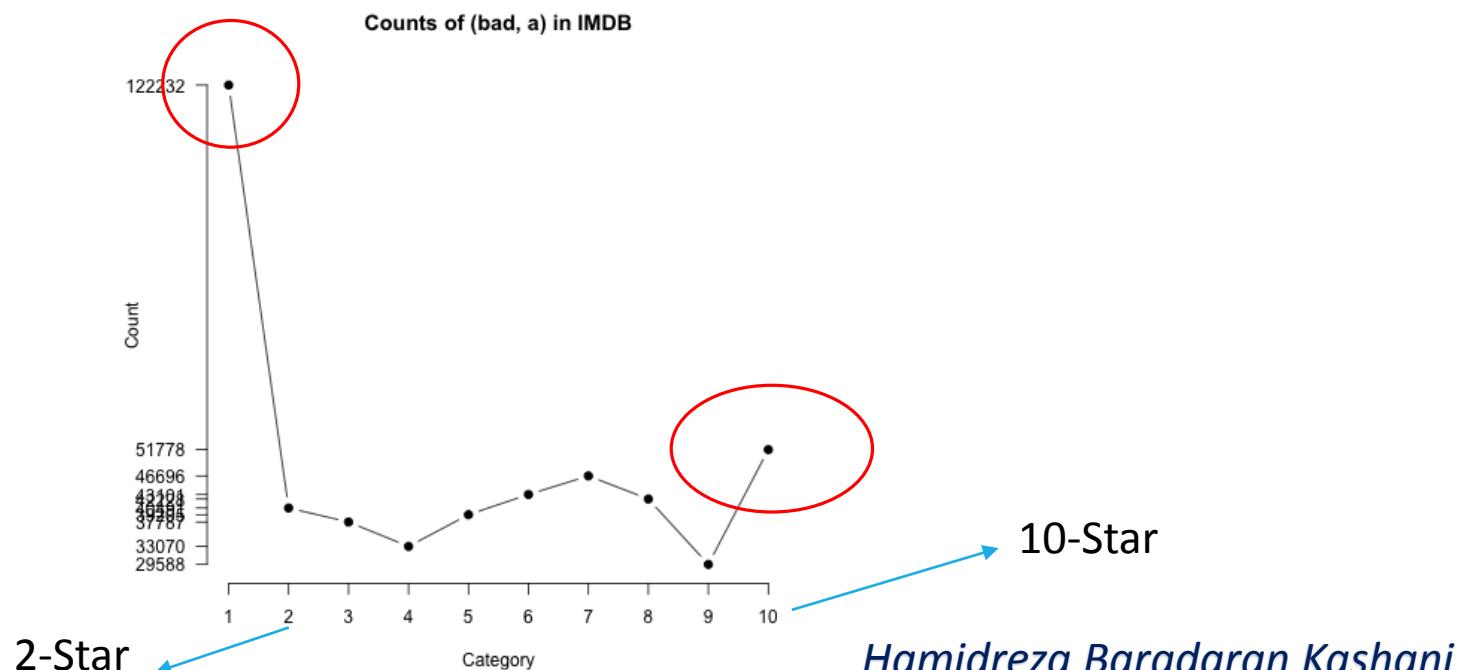
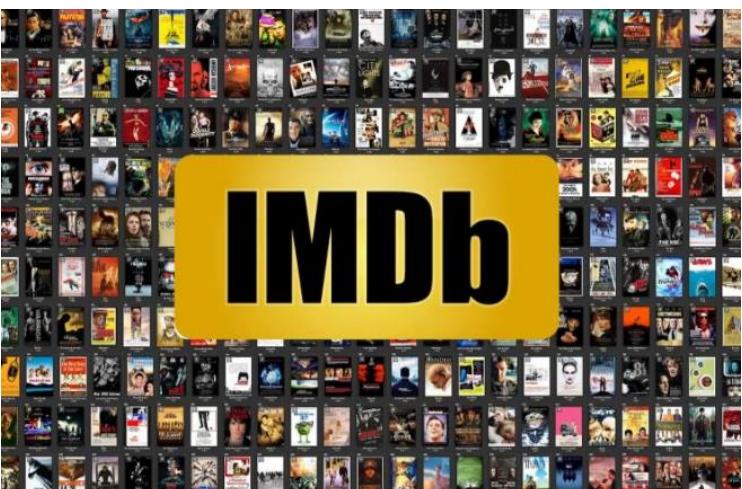
	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
General Inquirer			520/2306 (23%)	1/204 (0.5%)
SentiWordNet				174/694 (25%)
LIWC				

Hamidreza Baradaran Kashani



بررسی قطبیت کلمات در IMDB

- ❖ احتمال وقوع یک کلمه در هر یک از کلاس های احساس چقدر است؟
- ❖ داده IMDB مربوط به امتیازاتی است که بینندگان به هر فیلم داده اند (از یک ستاره تا ۱۰ ستاره)
- ❖ استفاده از **تعداد وقوع خام یک کلمه** روش مناسبی برای بدست آوردن قطبیت یک کلمه نیست.





بررسی قطبیت کلمات در IMDB

Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

❖ استفاده از معیار دیگری برای محاسبه قطبیت یک کلمه در یک مجموعه داده

likelihood



$$P(w | c) = \frac{f(w, c)}{\sum_{w \in c} f(w, c)}$$

Scaled likelihood

$$\frac{P(w | c)}{P(w)}$$

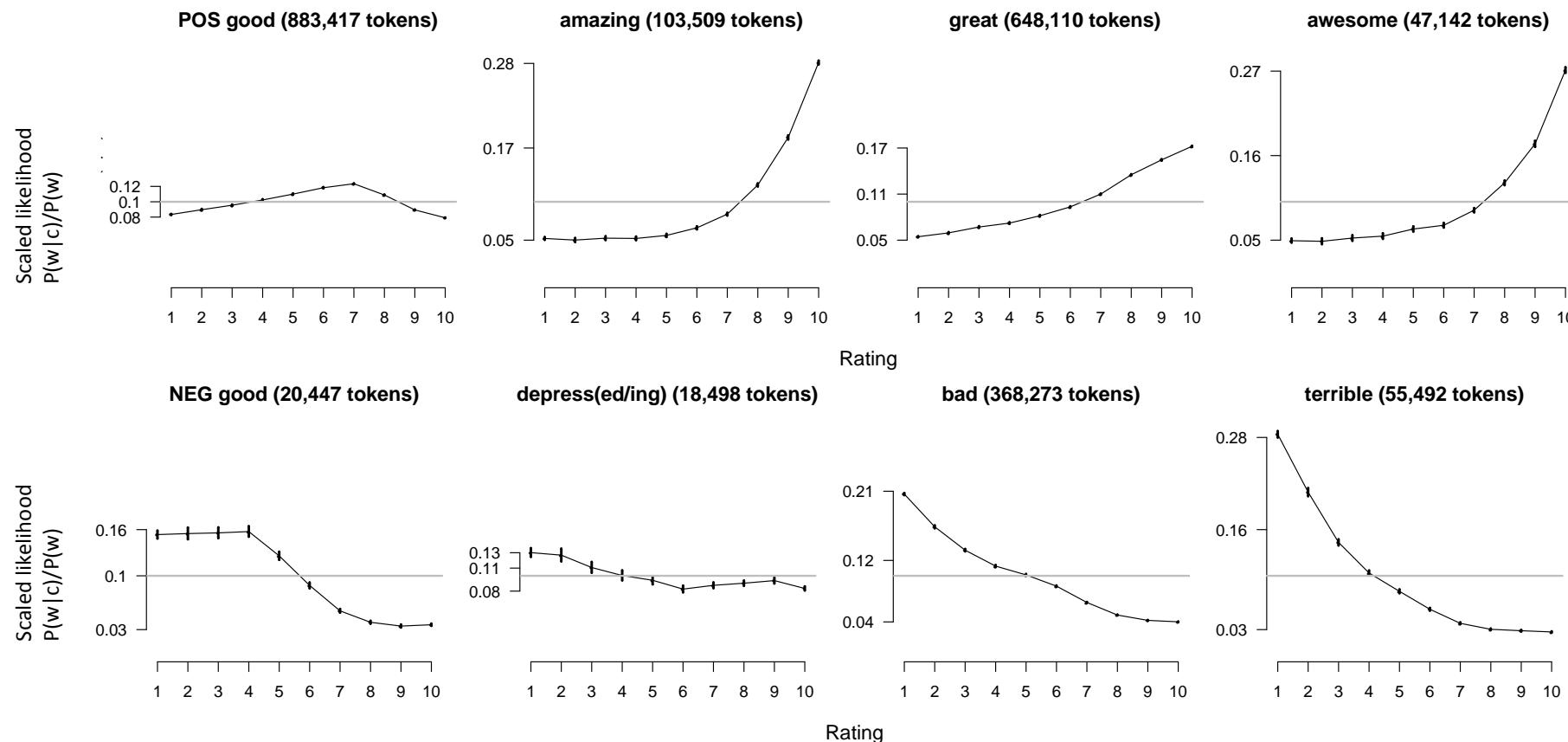
❖ معیار **Scaled likelihood** : نرمالیزه کردن معیار **likelihood** با فراوانی رخداد کلمات

Hamidreza Baradaran Kashani



بررسی قطبیت کلمات در IMDB

Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.



نقیض در قطبیت کلمات

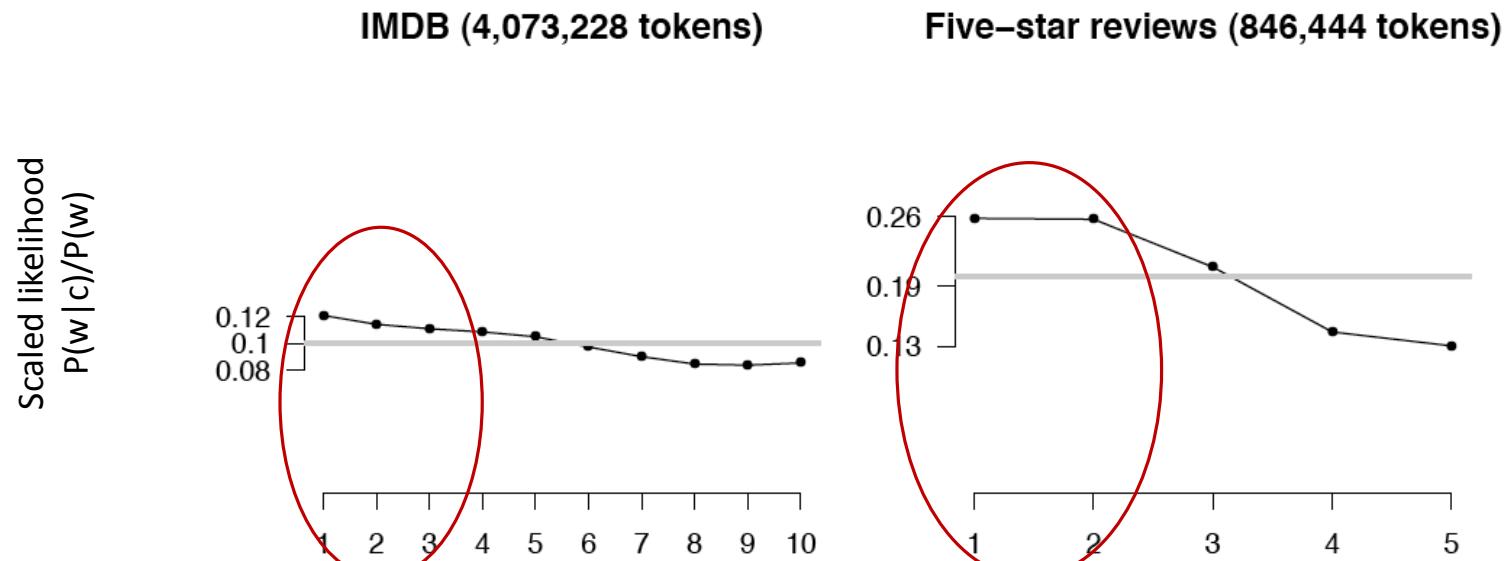


Potts, Christopher. 2011. On the negativity of negation. SALT 20, 636-659.

❖ آیا کلمات نقیض کننده مانند **not**, **no** و ... تنها در متون منفی دیده می شود؟

❖ Potts experiment:

❖ تعداد وقوع نقیض کننده ها مانند (**not**, **n't**, **no**, **never**) را در نظرات مجموعه داده ها شمارش کرده و مقدار **scaled likelihood** را برای درجات قطبیت مختلف محاسبه کرده است.



کلمات نقیض کننده
بیشتر در متون با
قطبیت منفی رخ
داده است



یادگیری (ساخت) یک واژه نامه احساسات

Hamidreza Baradaran Kashani



ساخت واژه نامه ها با روش نیمه نظارتی

- ❖ با وجود آنکه واژه نامه های احساس مختلفی طراحی شده اند، اما در بعضی موارد نیاز داریم این واژه نامه ها را توسعه دهیم و یا بخصوص برای یک دامنه کاربردی خاص واژه نامه های جدیدی بسازیم.
- ❖ رویکردهای نیمه نظارتی (**Semisupervised**) می توانند برای این منظور استفاده شوند.
- ❖ بطور کلی در رویکردهای نیمه نظارتی (**Semisupervised**) ما تعداد محدودی داده برچسب دار به همراه بخش قابل توجهی داده بدون برچسب داریم.
- ❖ مثلاً تعداد محدودی کلمه داریم که برچسب مثبت و منفی برای احساس آنها در دسترس هست.
- ❖ حال از این داده محدود استفاده شده و با مثلاً رویکردهای مختلف مثلاً یادگیری ماشین سایر کلمات برچسب گذاری شده و یک واژه نامه کامل ایجاد می شود.



روش Hatzivassiloglou and McKeown

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

ایده اصلی:

صفاتی که با کلمه "and" به هم متصل شده اند دارای قطبیت یکسانی هستند.

- Fair **and** legitimate, corrupt **and** brutal
- *fair **and** brutal, *corrupt **and** legitimate

صفاتی که با کلمه "but" به هم متصل شوند دارای قطبیت متفاوتی هستند.

- fair **but** brutal
- گران اما زیبا



روش گام اول – Hatzivassiloglou and McKeown

برای شروع ابتدا ۱۳۳۶ صفت مثبت و منفی را تحت عنوان seed set جمع آوری کردند. ♦

- **657 positive**

- adequate central clever famous intelligent remarkable reputed sensitive slender thriving...

- **679 negative**

- contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting...



روش گام دوم – Hatzivassiloglou and McKeown

❖ گسترش مجموعه اولیه با استفاده از الگوی صفات به هم پیوسته با **and** و **but**

Google "was nice and"

Nice location in Porto and the front desk staff **was nice and helpful** ...
www.tripadvisor.com>ShowUserReviews-g189180-d206904-r12068... +1

nice, helpful

Mercure Porto Centro: Nice location in Porto and the front desk staff **was nice and helpful** - See traveler reviews, 77 candid photos, and great deals for Porto, ...

If a girl **was nice and classy**, but had some vibrant purple dye in ...
answers.yahoo.com › Home › All Categories › Beauty & Style › Hair +1

nice, classy

4 answers - Sep 21

Question: Your personal opinion or what you think other people's opinions might ...

Top answer: I think she would be cool and confident like katy perry :)

Hamidreza Baradaran Kashani

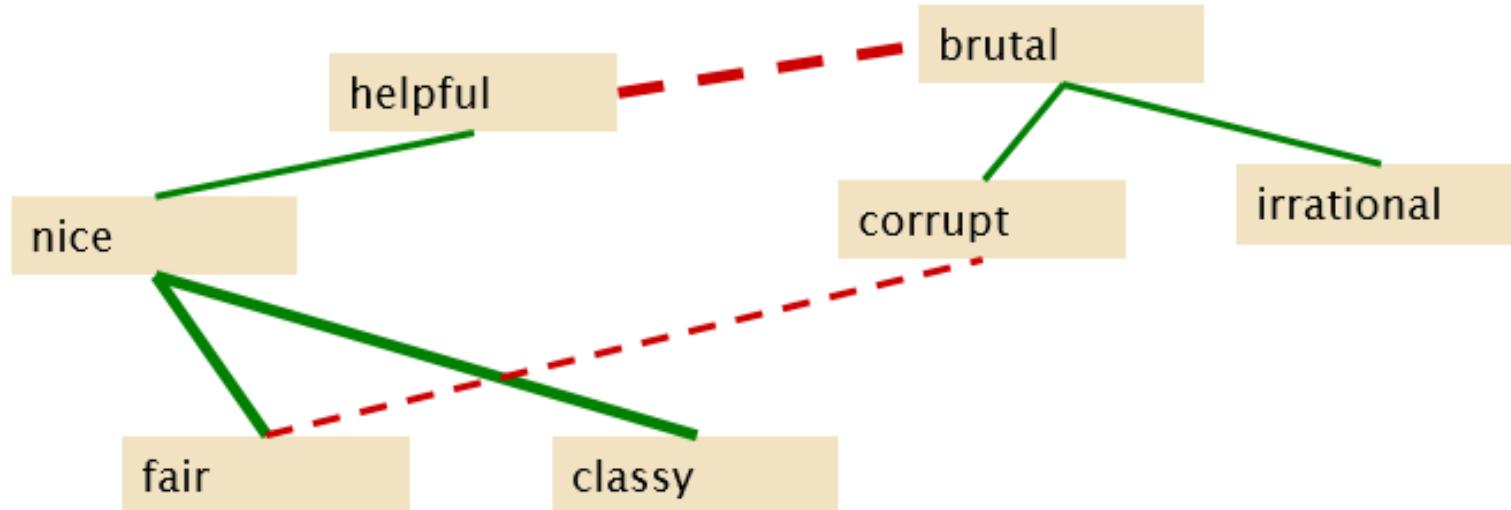


روش ۳ - گام سوم Hatzivassiloglou and McKeown

- ❖ ساخت یک گراف از تمام صفات
- ❖ نودهای گراف در واقع همان صفات اولیه و گسترش یافته هستند.
- ❖ مقدار لبه ها (edges) در گراف بیانگر "شباهت قطبیت" دو کلمه است که می تواند توسط یک دسته بند محاسبه می شود.
- مثلا بر روی پیکره یادگیری نسبت تعداد دفعاتی که دو تا صفت با and (یا but) آمده اند نسبت به کل تعداد دفعات به عنوان میزان شباهت آن دو در نظر گرفته شود.
- مثلا آموزش یک دسته بند که آن صفات که با and آمده اند را به کلاس ۱ و آن صفاتی که با but آمده اند را به کلاس ۱ - تخصیص دهد.



روش گام سوم – Hatzivassiloglou and McKeown



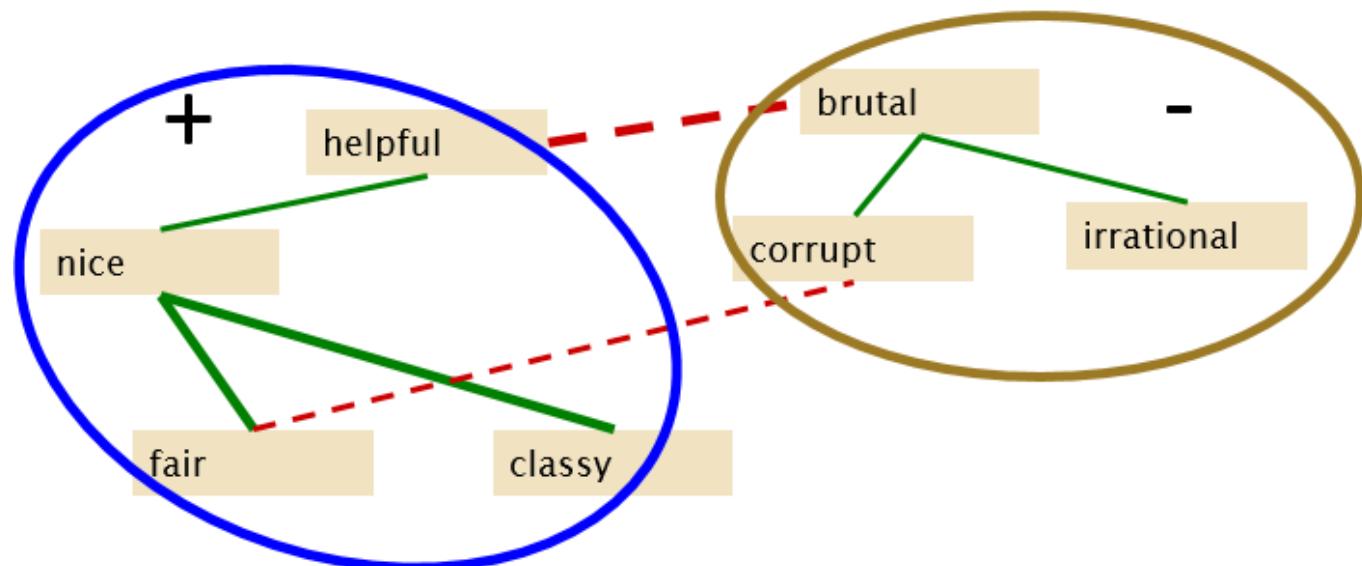
- برای مثال در اینجا شباهت دو کلمه **nice** و **fair** بیشتر است از شباهت دو کلمه **nice** با **helpful**
- یا مثلا **corrupt** و **fair** با یکدیگر تفاوت دارند به بیان دیگر بیشتر دفعات آنها با **but** به همدیگر متصل شده بودند تا and
- مثلا **nice** و **classy** شباهت زیادی به همدیگر دارند (خط سبز رنگ ضخیم) (در بیشتر موارد آنها با **and** به یکدیگر متصل شده اند)



روش چهارم – Hatzivassiloglou and McKeown

❖ خوشه بندی گراف به دو خوشه

- ❖ برای تمام صفاتی که داخل خوشه اول قرار گرفتند (خوشه ای که بیشتر صفات آن در دسته بندی مرحله اول و دوم مثبت بودند) را در نهایت **قطبیت مثبت** در نظر گرفت.
- ❖ برای تمام صفاتی که داخل خوشه دوم قرار گرفتند (خوشه ای که بیشتر صفات آن در دسته بندی مرحله اول و دوم منفی بودند) را در نهایت **قطبیت منفی** در نظر گرفت.



Hamidreza Baradaran Kashani



Positive

- bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty...

Negative

- ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful...

این خروجی حاوی خطاهایی نیز می باشد !



Positive

- bold decisive **disturbing** generous good honest important large mature patient peaceful positive proud sound stimulating straightforward **strange** talented vigorous witty...

Negative

- ambiguous **cautious** cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor **outspoken pleasant** reckless risky selfish tedious unsupported vulnerable wasteful...



روش Turney

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

❖ ایده اصلی:

❖ استخراج قطبیت برای عبارات به جای کلمات با استفاده از یک رویکرد نیمه نظارتی

❖ مراحل روش

- استخراج عبارات از یک پیکره یادگیری
- یادگیری و تخمین قطبیت هر عبارت
- تخمین قطبیت هر نظر (review) بر حسب قطبیت عبارات دیده شده در آن



استخراج عبارات دو کلمه‌ای شامل صفت

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything



محاسبه قطبیت یک عبارت

❖ ایده اصلی:

- ❖ کلمات مثبت در همسایگی کلمات مثبت بیشتر ظاهر می شوند.
 - مثلا عبارات مثبت معمولاً بیشتر با کلمه **excellent** رخ می دهند.

- ❖ کلمات منفی در همسایگی کلمات منفی بیشتر می آیند.
 - مثلا عبارات منفی معمولاً بیشتر با کلمه **poor** رخ می دهند.

❖ سوال:

چگونه می توان همسایگی کلمات یا رخداد همزمان (**co-occurrence**) کلمات را محاسبه کرد؟



Pointwise Mutual Information

❖ Mutual Information

❖ اطلاعات متقابل بین دو متغیر تصادفی X و Y

$$I(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

❖ Pointwise Mutual Information (PMI)

❖ از بین حالت‌هایی که X و Y اتفاق می‌افتد، چه تعداد از آنها با هم اتفاق می‌افتد؟

$$\text{PMI}(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$



Pointwise Mutual Information

❖ Pointwise Mutual Information (PMI)

$$\text{PMI}(X, Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

❖ بین دو کلمه PMI

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$



- ❑ $P(\text{word})$ estimated by **hits (word) / N**
- ❑ $P(\text{word}_1, \text{word}_2)$ by **hits (word1 NEAR word2) / N²**

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{P(\text{word}_1, \text{word}_2)}{P(\text{word}_1)P(\text{word}_2)}$$

$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \frac{\text{hits}(\text{word}_1 \text{ NEAR } \text{word}_2)}{\text{hits}(\text{word}_1)\text{hits}(\text{word}_2)}$$



یک عبارت بیشتر با poor رخ داده است یا excellent

$$\text{Polarity}(\text{phrase}) = \text{PMI}(\text{phrase}, \text{"excellent"}) - \text{PMI}(\text{phrase}, \text{"poor"})$$

$$= \log_2 \frac{\text{hits}(\text{phrase NEAR "excellent"})}{\text{hits}(\text{phrase})\text{hits}(\text{"excellent"})} - \log_2 \frac{\text{hits}(\text{phrase NEAR "poor"})}{\text{hits}(\text{phrase})\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(\text{phrase NEAR "excellent"})}{\text{hits}(\text{phrase})\text{hits}(\text{"excellent"})} \frac{\text{hits}(\text{phrase})\text{hits}(\text{"poor"})}{\text{hits}(\text{phrase NEAR "poor"})}$$

$$= \log_2 \frac{\ddot{\text{e}}^{\text{hits}(\text{phrase NEAR "excellent"})\text{hits}(\text{"poor"})} \ddot{\theta}}{\ddot{\text{e}}^{\text{hits}(\text{phrase NEAR "poor"})\text{hits}(\text{"excellent"})} \ddot{\theta}}$$



مثال‌هایی از عبارات مثبت و منفی

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
...		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
<i>Average</i>		0.32

Hamidreza Baradaran Kashani



مثال‌هایی از عبارات مثبت و منفی

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
...		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
Average		-1.2

Hamidreza Baradaran Kashani



نتائج الگوریتم Turney

۴۱۰ نظر را بررسی کرده است:

۱۷۰ نظر منفی (٪.۴۱)

۲۴۰ نظر مثبت (٪.۵۹)

٪.۵۹ : (Majority class) روش پایه

٪.۷۴ : Turney الگوریتم

❖ در روش پیشنهادی Turney از عبارات به جای کلمات استفاده می شود.

❖ می توان عبارات مثبت و منفی برای یک دامنه خاص را استخراج کرد (domain-specific)

استفاده از وردنت

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. COLING 2004
M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, 2004

- ❖ وردنت یک تزاروس است که شامل ارتباطات معنایی مختلف بین کلمات است مثلا:
 - ارتباط هم معنی بودن (**Synonymy**) و متضاد بودن (**Antonym**)

- ❖ روش ساخت واژه نامه:
 - ❖ از تعداد کلمه اولیه مثبت (مثلا **good**) و منفی (مثلا **terrible**) شروع می کنیم.
 - ❖ ساخت مجموعه مثبت: هم معنی های کلمات مثبت و متضادهای کلمه منفی
 - ❖ ساخت مجموعه منفی: هم معنی های کلمات منفی و متضادهای کلمه مثبت
 - مثلا اضافه کردن کلمه **awful** (هم معنی **terrible**) و کلمه **evil** (متضاد کلمه **good**) به مجموعه منفی
 - ❖ ادامه مراحل فوق تا زمانی که کلمه جدیدی تولید نشود.
 - ❖ فیلتر کردن کلمات نامرتب



خلاصه روش های ساخت واژه نامه

✓ امکان ساخت واژه نامه های وابسته به یک دامنه خاص

- امکان وجود احساس متفاوت برای یک کلمه در یک دامنه خاص یا عدم وجود احساس برای یک کلمه در واژه نامه های عام
 - کلمه "بزرگ" در حوزه اتاق های یک هتل حس مثبت دارد اما برای یک دوربین می تواند حس منفی داشته باشد.
 - در حوزه بانکداری عبارت "پرداخت مستقیم" می تواند قطبیت مثبت داشته باشد (در واژه نامه های عام وجود ندارد)

❖ رویکرد کلی روش ها

- استفاده از یک سری کلمات اولیه مناسب و دقیق (seed words) مثل **poor** و **good**
- پیدا کردن کلمات با قطبیت مشابه به این کلمات با روش های مختلف:
 - استفاده از **but** و **and**
 - کلمات مجاور کلمات مثبت و منفی اولیه در اسناد یکسان و مشابه
 - استفاده از روابط هم معنی و تضاد موجود در وردنت



سایر موارد



یافتن احساس مبتنی بر جنبه

✓ در برخی موارد لزوماً یک جمله دارای یک قطبیت نیست، مثلاً:

- The **food** was great but the **service** was awful
- مثلاً در جمله بالا نسبت به مشخصه یا جنبه "غذا" قطبیت مثبت است اما نسبت به جنبه "سریس دهی" قطبیت منفی است.

❖ نام های دیگر برای جنبه (**aspect**)

○ ویژگی یا مشخصه (**attribute**)

○ هدف (**target**)



M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In Proceedings of KDD.

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop.

یافتن احساس مبتنی بر جنبه

❖ نحوه یافتن جنبه ها در یک متن (استفاده از قواعد)

- عباراتی که فراوان در یک مجموعه از نظرات تکرار شده اند (بخصوص اسم ها)
- در انگلیسی (فارسی) کلماتی که پس (قبل) از کلمات دارای قطبیت در جمله می آیند، معمولاً یک aspect می باشند.

▪ "فیلم زیبایی بود" یا "... delicious food ..."

○ استفاده از لیست جنبه های استخراج شده از دامنه های مختلف

Casino	casino, buffet, pool, resort, beds
Children's Barber	haircut, job, experience, kids
Greek Restaurant	food, wine, service, appetizer, lamb
Department Store	selection, department, sales, shop, clothing



یافتن احساس مبتنی بر جنبه

- ✓ ممکن است جنبه ها بطور واضح در جمله ظاهر نشوند، مثلا:
 - ✓ مثلا در جمله "خانه خیلی گران است" در اینجا جنبه مورد نظر "قیمت" است.
 - ✓ البته در دامنه هایی مثل هتل ها و رستوران ها راحت تر می توان جنبه ها را یافت.

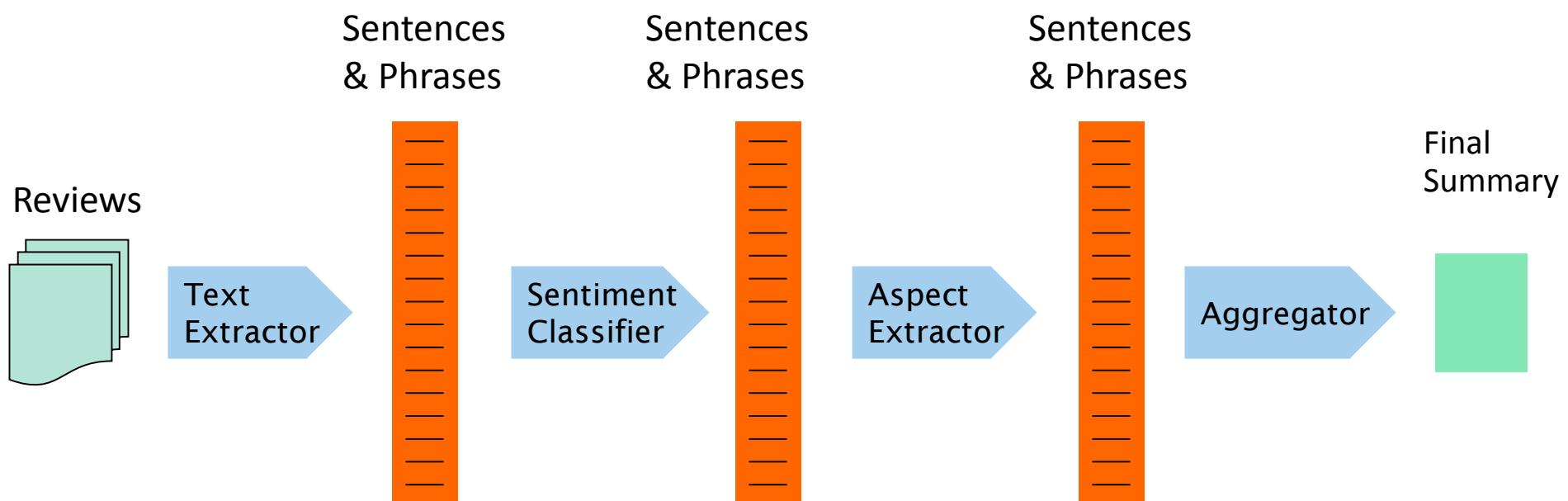
❖ رویکرد یادگیری بانظارت

- در یک مجموعه اولیه از نظرات (مثلا در رابطه با رستوران)، جنبه ها را مشخص و بصورت دستی برچسب گذاری کنیم (مثلا قیمت، کیفیت، سرویس دهی و غیره)
- سپس یک دسته بند آموزش دهیم که نوع aspect درون یک جمله را تخمین بزند، مثلا یک دسته بند چهار کلاسه با کلاس های زیر:
- food, décor, service, value, NONE



یافتن احساس مبتنی بر جنبه

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008.
Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop





Rooms (3/5 stars, 41 comments)

- (+) The room was clean and everything worked fine – even the water pressure ...
- (+) We went because of the free room and was pleasantly pleased ...
- (-) ...the worst hotel I had ever stayed at ...

Service (3/5 stars, 31 comments)

- (+) Upon checking out another couple was checking early due to a problem ...
- (+) Every single hotel staff member treated us great and answered every ...
- (-) The food is cold and the service gives new meaning to SLOW.

Dining (3/5 stars, 18 comments)

- (+) our favorite place to stay in biloxi.the food is great also the service ...
- (+) Offer of free buffet for joining the Play



تعداد نمونه های نامتوازن در کلاس ها

- ❖ در حالت کلی فرض می شود که تعداد نمونه های کلاس های مختلف (مثلاً مثبت و منفی) تقریباً یکسان هستند. اما در عمل و داده های واقعی این گونه نیست.
 - در این موارد باید از معیار ارزیابی **Accuracy** استفاده کرد.
 - معیار **F-score** برای حالت نامتوازن بودن کلاس ها معیار مناسبی است.
 - عدم توازن (شدید) کلاس ها تاثیر منفی در عملکرد کلاسیفایر دارد.

❖ دو روش رایج برای حل مساله نامتوازن بودن

- ✓ نمونه برداری از سمپل های کلاس بیشتر (جهت بالانس شدن نمونه ها با تعداد نمونه ها در کلاس کوچکتر)
- ✓ تغییر در تابع هزینه
- ✓ مثلاً خطاهای دسته بندی نمونه های مربوط به مجموعه کوچکتر با وزن بیشتری در تابع هزینه لحاظ شوند (تمرکز بیشتر تابع هزینه بر روی نمونه های کلاس کوچکتر)



آنالیز احساسات، مساله دسته بندی یا رگرسیون؟

❖ آنالیز احساسات را تحت عنوان دسته بندی احساسات (Sentiment classification) نیز به کار می برند.



❖ بطور کلی یک مساله باینری در نظر گرفته می شود (ثبت در مقابل منفی)

❖ سه کلاس: ثبت، خنثی و منفی

❖ پنج کلاس: خیلی ثبت، ثبت، خنثی، منفی، خیلی منفی



❖ در سامانه های توصیه گر نیاز به پیش بینی امتیاز یک کاربر به یک محصول یا مشخصه خاص از آن است.

❖ کاربران امتیاز را بصورت ۵ ستاره (خیلی خوب) تا ۱ ستاره (خیلی ضعیف) می دهند.



آنالیز احساسات، مساله دسته بندی یا رگرسیون؟

❖ ۲ تا رویکرد برای حل مسائل آنالیز نظرات یا احساسات بیشتر از ۲ کلاس:

❖ راهکار اول) تبدیل به یک حالت باینری

- مثلا برای یک مساله ۵ کلاسی امتیازات بالای ۲.۵ به عنوان کلاس مثبت و امتیازات کمتر از ۲.۵ به عنوان کلاس منفی در نظر گرفته شوند.
- این روش دقیق نیست!!



آنالیز احساسات، مساله دسته بندی یا رگرسیون؟

❖ راه حل دوم و بهتر: استفاده از رگرسیون به جای دسته بندی

- کلاس های احساسات یا نظرات مستقل از **یکدیگر نیستند** بلکه با توجه به مقدارشان به یکدیگر مرتبط هستند.
- امتیاز ۵ یک واحد از ۴ بیشتر است، امتیاز ۴ یک واحد از ۳ بیشتر است، اما امتیاز ۴ سه واحد بیشتر از ۱ است
- اگر کاربری به فیلمی امتیاز ۵ دهد احتمال آنکه به فیلم بسیار مشابهی به فیلم اول، امتیاز ۴ دهد بیشتر است تا امتیاز ۲ (به بیان دیگر ۴ و ۵ به هم شبیه تر هستند و همین طور ۱ و ۲ به همدیگر)
- به بیان دیگر ترتیب بین خروجی های یک سیستم توصیه گر یا آنالیز نظرات داریم.

- متفاوت از یک مساله دسته بندی شی داخل تصویر (گربه، دوچرخه، سگ و ...)
- در این مساله ترتیب (ordering) بین کلاس ها نداریم.



تعداد کلاس های بیشتر از ۲ (مثلا ۵ یا ۷)

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. ACL, 115–124

- ❖ تبدیل به یک حالت باینری
- ❖ مثلا امتیازات بالای ۳.۵ به عنوان کلاس مثبت و امتیازات کمتر از ۳.۵ به عنوان کلاس منفی
- ❖ استفاده از رگرسیون به جای دسته بندی



MICROBLOG COMMENTS WITH EMOJIS.

emoji	sentiment	microblog comments
😢	positive	The clothes I ordered arrived and they look beautiful. 😊
	negative	My stomach hurts, I don't want to talk. 😢
😊	positive	It is a nice day, I feel good. 😊
	negative	I do not want to say anymore. 😊 😊 😊