

Naive Bayes and Text Classification

Lecture 6

Text Categorization

- text categorization, the task of assigning a label or categorization category to an entire text or document.

Is this spam?

Subject: Important notice!

From: Stanford University <newsforum@stanford.edu>

Date: October 28, 2011 12:34:16 PM PDT

To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochinchina; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

Positive or negative movie review?



- 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.

What is the subject of this article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

What is Sentiment Analysis ?

- The positive or negative orientation that a writer expresses toward some object

Sentiment analysis has many other names

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

Why sentiment analysis?

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

Positive or negative movie review?



- 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.

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 shi

Reviews

Summary - Based on 377 reviews



What people are saying

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

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)



Most mentioned

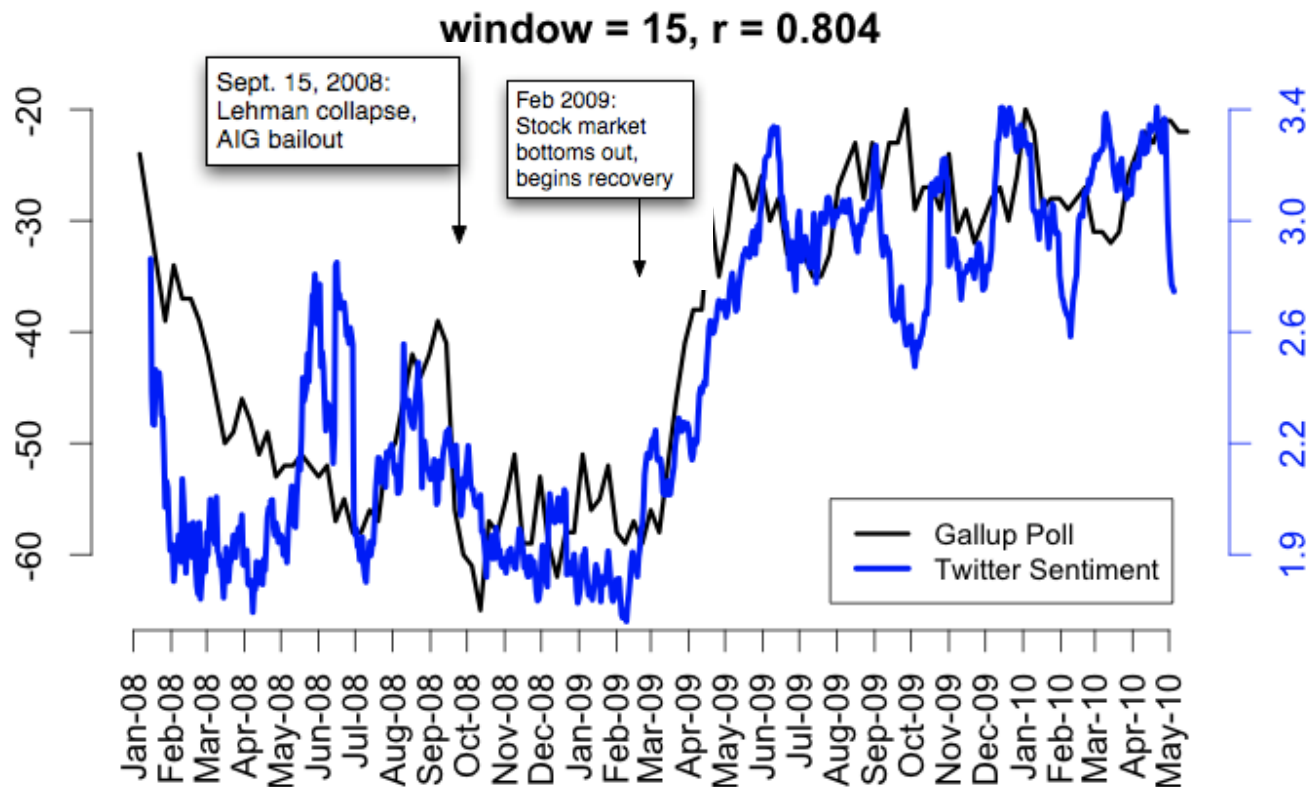


Show reviews by source

Best Buy (140)
CNET (5)
Amazon.com (3)

Twitter sentiment versus Gallup Poll of Consumer Confidence

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



Target Sentiment on Twitter

- [Twitter Sentiment App](#)

- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision

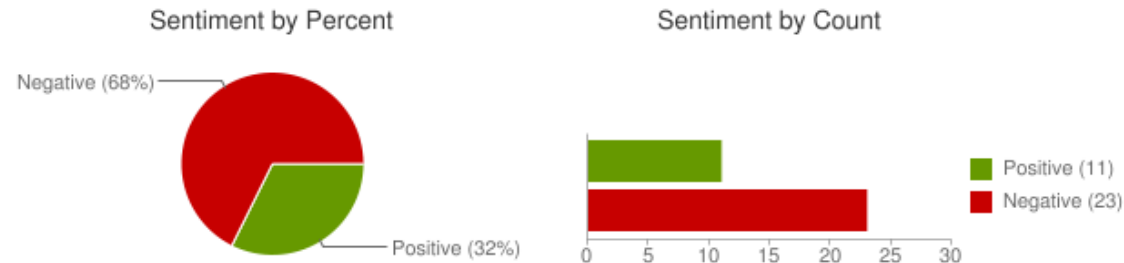
Type in a word and we'll highlight the good and the bad

"united airlines"

Search

[Save this search](#)

Sentiment analysis for "united airlines"



[jljacobson](#): 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](#)

Sentiment Analysis

- Sentiment analysis is the detection of **attitudes**
“enduring, affectively colored beliefs, dispositions towards objects or persons”
 1. **Holder (source)** of attitude
 2. **Target (aspect)** of attitude
 3. **Type** of attitude
 - From a set of types
 - *Like, love, hate, value, desire, etc.*
 - Or (more commonly) simple weighted **polarity**:
 - *positive, negative, neutral, together with strength*
 4. **Text** containing the attitude
 - Sentence or entire document

Sentiment Analysis

- Simplest task:
 - Is the attitude of this text positive or negative?
- More complex:
 - Rank the attitude of this text from 1 to 5
- Advanced:
 - Detect the target, source, or complex attitude types

Sentiment Analysis

- Simplest task:
 - Is the attitude of this text positive or negative?
- More complex:
 - Rank the attitude of this text from 1 to 5
- Advanced:
 - Detect the target, source, or complex attitude types

Text Classification: definition

- *Input*:
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods:

Hand-coded rules

- Rules based on combinations of words or other features
 - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
 - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- *Input:*
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
 - A training set of m hand-labeled documents
 $(d_1, c_1), \dots, (d_m, c_m)$
- *Output:*
 - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
 - Naïve Bayes
 - Logistic regression
 - Support-vector machines
 - k-Nearest Neighbors
 - ...

Sentiment Analysis

Naïve Bayes Algorithm

Sentiment Classification in Movie Reviews

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

- Polarity detection:
 - Is an IMDB movie review positive or negative?
- Data: *Polarity Data 2.0*:
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data>

IMDB data in the Pang and Lee database



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 .

Baseline Algorithm

- Tokenization
- Feature Extraction
- Classification using different classifiers

Sentiment Tokenization Issues

- Deal with HTML and XML markup
- Twitter mark-up (names, hash tags)
- Capitalization (preserve for

words in all caps)

- Phone numbers, dates
- Emoticons
- Useful code:

Potts emoticons

[<>]?	# optional hat/brow
[:;=8]	# eyes
[\-o*\']?	# optional nose
[\)\]\]\(\[dDpP/\:\}\{\@\\ \\]	# mouth
	#### reverse orientation
[\)\]\]\(\[dDpP/\:\}\{\@\\ \\]	# mouth
[\-o*\']?	# optional nose
[:;=8]	# eyes
[<>]?	# optional hat/brow

- [Christopher Potts sentiment tokenizer](#)
- [Brendan O'Connor twitter tokenizer](#)

Extracting Features for Sentiment Classification

- How to handle negation
 - I **didn't** like this movie
 - vs
 - I really like this movie
- Which words to use?
 - Only adjectives
 - All words
 - All words turns out to work better, at least on this data

Negation

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.

Add NOT_ to every word between negation and following punctuation:

didn't like this movie , but I



didn't NOT_like NOT_this NOT_movie but I

Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Relies on very simple representation of document
 - Bag of words

The bag of words representation

Y(

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

)=C





The bag of words representation

$Y($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

$) = C$

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

$$P(c | d) = \frac{P(d | c)P(c)}{P(d)}$$

$$P(\text{Pos} | d) = P(d | \text{Pos}) P(\text{Pos}) / P(d)$$

$$P(\text{Neg} | d) = P(d | \text{Neg}) P(\text{Neg}) / P(d)$$

Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Dropping the denominator

Naïve Bayes Classifier

Prior
Probability of
training class

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

Document d
represented as
features x1..xn

Multinomial Naïve Bayes

Independence Assumptions

$$P(x_1, x_2, \square, x_n | c)$$

- **Bag of Words assumption:** Assume position doesn't matter
- **Conditional Independence:** Assume the feature probabilities $P(x_i | c_j)$ are independent given the class c .

$$P(x_1, \square, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \cdot P(x_3 | c) \cdot \dots \cdot P(x_n | c)$$

Multinomial Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c) P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x | c)$$

Applying Multinomial Naive Bayes Classifiers to Text Classification

positions \leftarrow all word positions in test document

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

Learning the Multinomial Naïve Bayes Model 13.3

- Maximum likelihood estimates
 - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive**?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{\text{count}(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0$$

$$c_{MAP} = \operatorname{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Naïve Bayes with Laplace (add-1) smoothing

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

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

Text Classification and Naïve Bayes

Multinomial Naïve Bayes: A Worked Example

Multinomial Naïve Bayes

$$= \underset{c \in C}{\operatorname{argmax}} P(d | c) P(c)$$

	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	?

$$\hat{P}(c) = \frac{N_c}{N}$$

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

	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	?

Priors:

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

Choosing a class:

$$P(c|d5) \propto \frac{3}{4} * \left(\frac{3}{7}\right)^3 * \frac{1}{14} * \frac{1}{14} \approx 0.0003$$

Conditional Probabilities:

$$P(\text{Chinese}|c) = (5+1) / (8+6) = 6/14 = 3/7$$

$$P(\text{Tokyo}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Japan}|c) = (0+1) / (8+6) = 1/14$$

$$P(\text{Chinese}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Tokyo}|j) = (1+1) / (3+6) = 2/9$$

$$P(\text{Japan}|j) = (1+1) / (3+6) = 2/9$$

$$P(j|d5) \propto \frac{1}{4} * \left(\frac{2}{9}\right)^3 * \frac{2}{9} * \frac{2}{9} \approx 0.0001$$

Underflow Prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow.
- Since $\log(xy) = \log(x) + \log(y)$
 - Better to sum logs of probabilities instead of multiplying probabilities.
- Class with highest un-normalized log probability score is still most probable.

$$c_{NB} = \operatorname{argmax}_{c_j \in C} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j)$$

- Model is now just max of sum of weights

Binarized (Boolean feature) Multinomial Naïve Bayes

- Intuition:
 - For sentiment (and probably for other text classification domains)
 - Word occurrence may matter more than word frequency
 - The occurrence of the word *fantastic* tells us a lot
 - The fact that it occurs 5 times may not tell us much more.
 - Boolean Multinomial Naïve Bayes
 - Clips all the word counts in each document at 1

Boolean Multinomial Naïve Bayes on a test document d

- First remove all duplicate words from d
- Then compute NB using the same equation:

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

Normal vs. Boolean Multinomial NB

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	?

Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Very good in domains with many equally important features
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification

Reference

- Speech and Language Processing By [Dan Jurafsky](#) and [James H. Martin](#) (3rd Edition)

Chapter 4 Naive Bayes and Sentiment
Classification

<https://web.stanford.edu/~jurafsky/slp3/4.pdf>