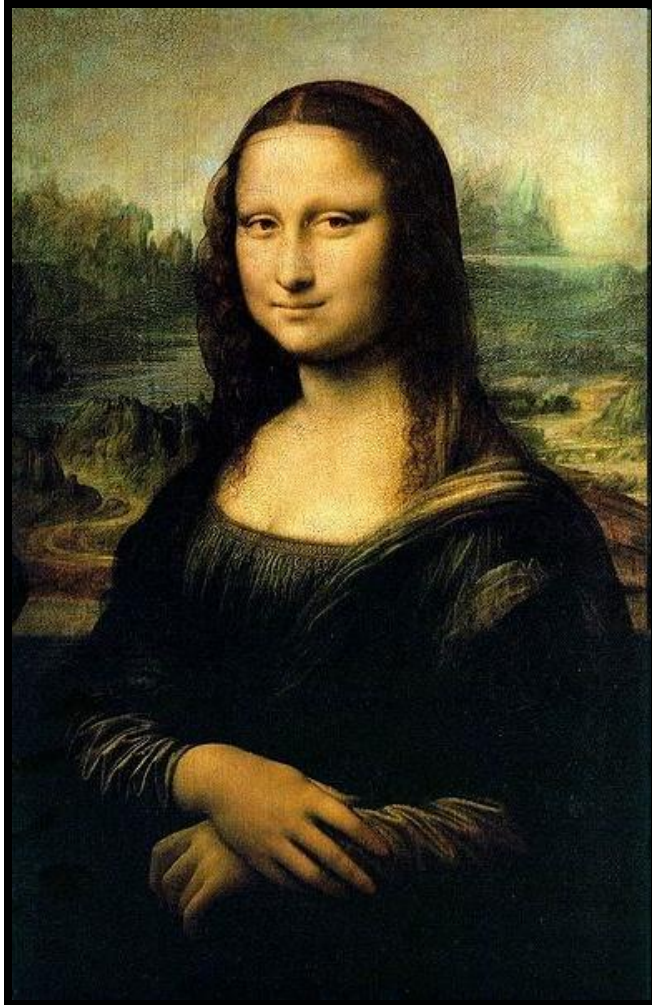


# Sentiment Analysis & Opinion Mining

Lecture One: March 1, 2011

Aditya M Joshi  
M Tech3, CSE  
IIT Bombay  
{adityaj@cse.iitb.ac.in}



## Smile of Mona Lisa

Is she smiling at all?

**Is she happy?**

What is she smiling about?

**What is she happy about?**

**Mona Lisa**

**16<sup>th</sup> century**

Artist: Leonardo da Vinci

# Sentiment analysis (SA)

Task of tagging text with orientation of opinion

*This is a good movie.* 

*This is a bad movie.* 

*The movie is set in  
Australia.* 

Subjective

Objective

# Outline

## Lecture 1

Motivation & Introduction

Classifiers for SA

## Lecture 2

Approaches to SA

Applications

# Outline

## Lecture 1

### Motivation & Introduction

Challenges of SA: *Why SA is non-trivial*

Variants of SA: *What forms does it exist in?*

Opinion on the web: *Is doing SA really worth it?*

### Classifiers for SA

## Lecture 2

### Approaches to SA

### Applications

# Challenges of SA

- Domain dependent
- Sarcasm
- Thwarted expressions
- Negation
- Implicit polarity
- Time-bounded

*“This phone allows me to send  
SMS.”*

*“This phone has a touch-screen.”*

# Flavours of SA

- Subjective/Objective
- Emotion analysis
- SA with magnitude
- Entity-specific SA
- Feature-based SA
- Perspectivization

*“The Leftists were arrested yesterday by the police.”*

# Opinion on the Web

- Does web really contain sentiment-related information?
- Where?
- How much?
- What?





# User-generated content

- Web 2.0 empowers the user of the internet
- They are most likely to express their opinion there
- Temporal nature of UGC: 'Live Web'
- **Can SA tap it?**

# Where?

- Blogs
- Review websites
- Social networks
- **User conversations**

Conversations between  
users on one of the above

# How much?

- Size of blogosphere
  - Through the 'eyes' of the blog trackers
- Technorati : 112.8 million blogs (excluding 72.82 million blogs in Chinese as counted by a corresponding Chinese Center)
- A blog crawler could extract 88 million blog URLs from blogger.com alone
- 12,000 new weblogs daily

# How much opinion?

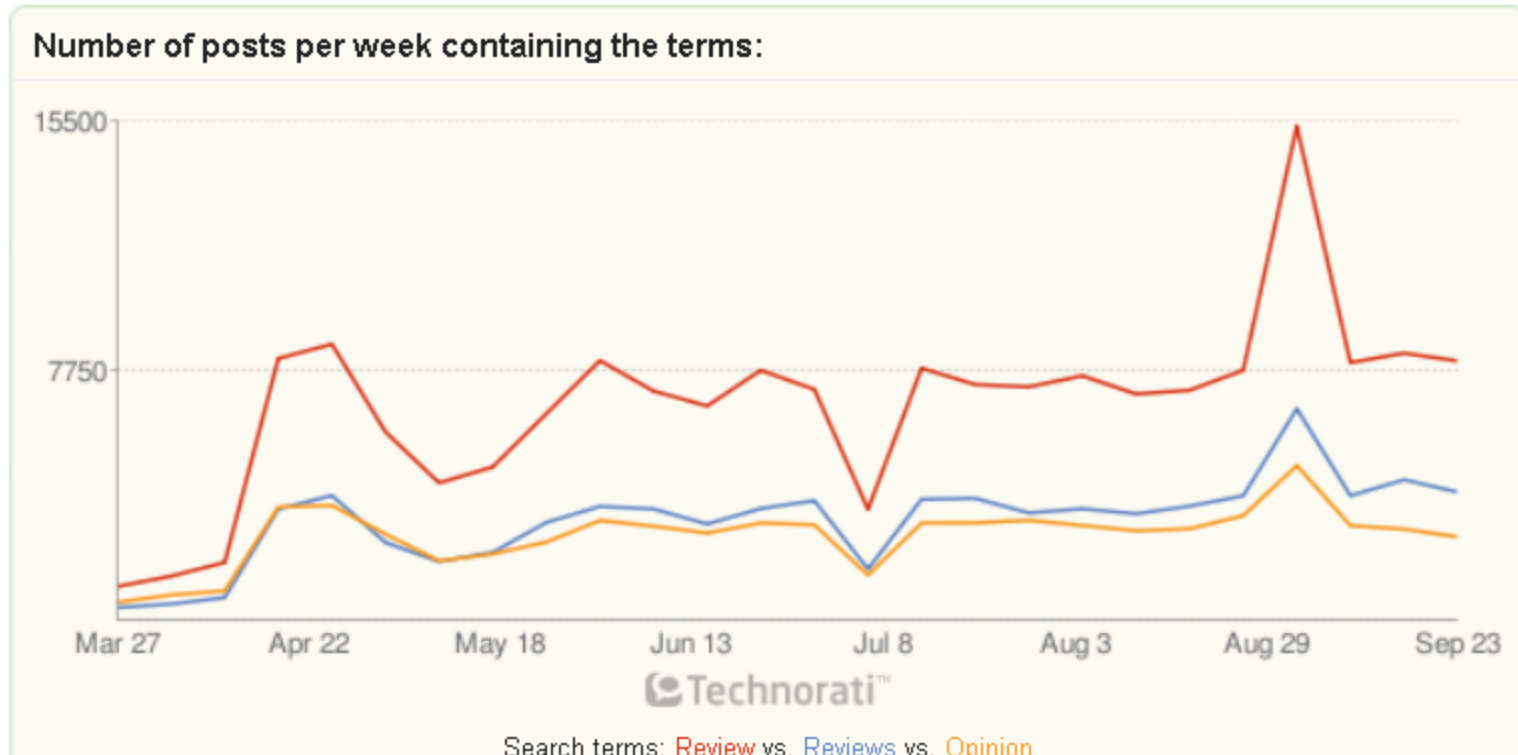
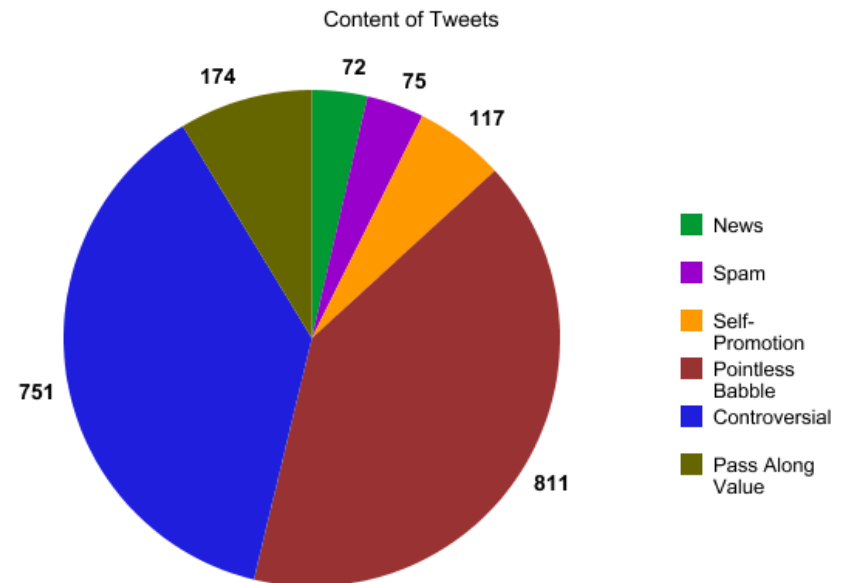


Chart created using : [www.technorati.com/chart/](http://www.technorati.com/chart/)

# How much?

- 12,22,20,617 unique visitors to facebook in December 2009
- Twitter:  
2,35,79,044



Kelly, Ryan, ed. (2009-08-12), "Twitter Study - August 2009" (PDF), Twitter Study Reveals Interesting Results About Usage, San Antonio, Texas: Pear Analytics. <http://www.pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf>

# What? Reviews

- [www.burrrp.com](http://www.burrrp.com)

Restaurant reviews (now, for a variety of 'lifestyle' products/services)

- [www.mouthshut.com](http://www.mouthshut.com)

A wide variety of reviews

- [www.justdial.com](http://www.justdial.com)

- [www.yelp.com](http://www.yelp.com)

Professionals: Well-formed

- [www.zagat.com](http://www.zagat.com)

User: More mistakes

- [www.bollywoodhungama.com](http://www.bollywoodhungama.com)

- [www.indya.com](http://www.indya.com)

Movie reviews by professional critics, users. Links to external reviews also present

# A typical Review website

The screenshot shows a review website for IIT Bombay. The page features a navigation bar with categories like Automobiles, Books, Computers, Electronics, Entertainment, Fashion, Food & Drinks, Health & Beauty, Personal Finance, Travel, and More. A search bar is located at the top right. The main content area displays a review titled "IIT - Bombay Review" by user "bunty007". The review includes a star rating of 5 stars, a recommendation of "No", and a "Great place to be in..." headline. The review text describes the user's experience at IIT Bombay. The page also features a sidebar with "About bunty007" information and a "Rate this review" section.

**MOUTHSHUT.COM** MouthShut India Select Your City Help | Invite Friends | Sign Up | Log In

Search: Type the name of product / memt Product Search

Automobiles | Books | Computers | Electronics | Entertainment | Fashion | Food & Drinks | Health & Beauty | Personal Finance | Travel | More


FREE SIGN UP CATEGORIES REVIEWS DIARIES PHOTOS POST A DIARY FRIENDS WRITE A REVIEW

Home > Education > Colleges: By State > Maharashtra Colleges > Bombay Colleges > IIT - Bombay > bunty007's review

Ads by Google Jobs Bangalore India Niit GIS LTD PG Diploma Pune Bangalore Girls Cheap Car Rentals

## IIT - Bombay Review

Product Details Current Review Review Comments Read All 5 Reviews Compare All Engineering Colleges Corporate Blog



**Great place to be in...**  
By: bunty007 | Jun 22, 2006 04:50 PM

Academic Programs:	██████████
Administration:	██████████
Extracurricular Programs:	██████████
Alumni Network:	██████████

Member's Rating: ★★★★★  
Member's Recommendation: **No**

Read **802** times  
Rated by **5** members

MouthShut Product Rating:  
★★★★★  
Recommended by **80%** members

Pros: **It's a good experience.**  
Cons: **It's not a suggestion to be in IITB only.**

[Write your own review](#) [SHARE THIS REVIEW](#)

IT Bombay, this name makes my blood boil...ofcourse in the positive sense. I spent my fabulous 4 years of life in there and I assure one and all of you this is the place to be in. Let's start with how it feels to be in there. For this I would like to describe my first week in there. This was the first time I was in Bombay...date: 16th July 2001. My heart was beating like anything when I reached the main gate and saw the logo on the gate with the motto Gyaanam Param Dhayaam. I was enticed by the very first look

### About bunty007

Name: Vivek Sharma  
...view complete profile

Reviews: 5  
Diary Posts: 0  
Trusted by: 11 members

[Trust this member](#) [Email this member](#)  
[Distrust this member](#) [Send a Gift](#)  
[Alert on new review by this member](#)

**Rate this review**  
(Earn 5 MS-Points™ by rating reviews)

Ads by Google Chevrolet Spark

# Sample Review 1

## (This, that and this)

FLY E300 is a good mobile which i purch is not familiar in Market as well known with almost all the features for a good would come around 19k Indian Ruppee

**'Touch screen' today signifies a positive feature. Will it be the same in the future?**

his Brand cheap of features

Touch Screen, good resolution, good talk time, 3.2Mega

**BUT BEWARE THAT THE CAMERA IS NOT THAT GOOD, TH ITS NOT AS GOOD AS MY PREVIOUS MOBILE SONY ERICS Pixel.**

**Comparing old products**

Sony ericsson was excellent with the feature of camera. So if anyone is thinking for Camera, please excuse. **This model of FLY is not apt for you..** Am fooled in this regard..

Audio is not bad, **infact better than Sony Ericsson K750**

FLY is not user friendly probably since we have just sta

**The confused conclusion**



# Sample Review 2

Hi,

I have Haier phone.. It was good when i was buing this phone.. But I invented A lot of bad features by this phone those are It's cost is low but Software is not good and Battery is very bad...,,Ther are no signals at out side of the city...,, People can't understand this type of software. There aren't features in this phone, Desig bad..So I'm not intrest this si it is good. They are giving me are also good.They are giving colour screen at display time it is also good because other phones aren't this type of feature.It is also low wait.

Lack of punctuation marks,  
Grammatical errors

Wait.. err.. Come again

# Sample Review 3

## (Subject-centric or not?)

I have this personal experience of using this cell phone. I bought it one and half years back. It had modern features that a normal cell phone has, and the look is excellent. I was very impressed by the design. I bought it for Rs. 8000. It was a gift for someone. It worked fine for first one month, and then started the series of multiple faults it has. First the speaker didnt work. I took it to the service centre (which is like a govt. office with no work). It took 15 days to repair the handset, moreover they charged me Rs. 500. Then after 15 days again the mike didnt work, then again same set of time was consumed for the repairs and it continued. Later the camera didnt work, the speakes were rubbish, it used to hang. It started restarting automatically. And the govt. office had staff which I doubt have any knoledge of cell phones??

These multiple faults continued for as long as one year, when the warranty period ended. In this period of time I spent a considerable amount on the petrol, a lot of time (as the service centre is a govt. office). And at last the phone is still working, but now it works as a paper weight. The company who produces such items must be sacked. I understand that it might be fault with one prticular handset, but the company itself never bothered for replacement and I have never seen such miserable cust service. For a comman man like me, Rs. 8000 is a big amount. And I spent almost the same amount to get it work, if any has a good suggestion and can gude me how to sue such companies, please guide.

For this the quality team is faulty, the cust service is really miserable and the worst condition of any organisation I have ever seen is with the service centre for Fly and Sony Ericson, (it's near Sancheti hospital, Pune). I dont have any thing else to say.

# Sample Review 4

(Good old sarcasm)

“I’ve seen movies where there was practically no plot besides explosion, explosion, catchphrase, explosion. I’ve even seen a movie where nothing happens. But *White on Rice* was new on me: a collection of really wonderful and appealing characters doing completely baffling and uncharacteristic things.”

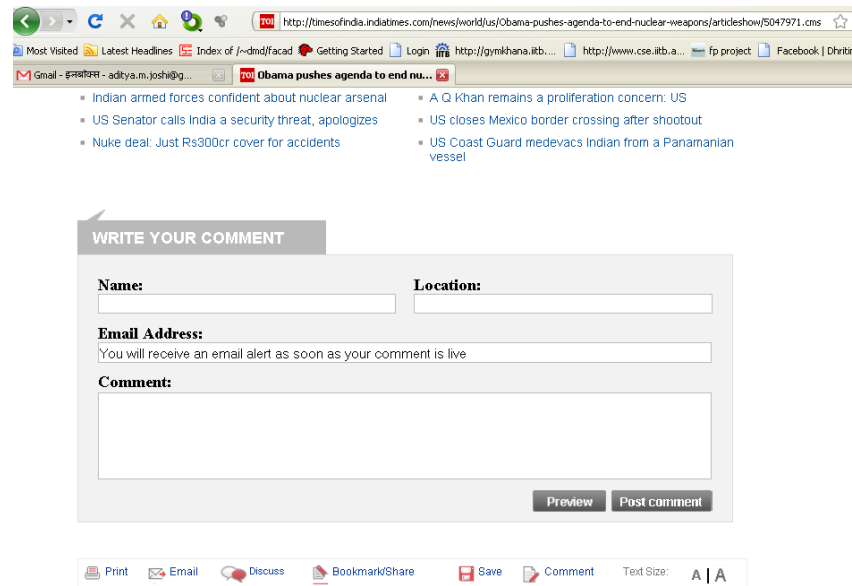
Review from: [www.pajiba.com](http://www.pajiba.com)

# What? Social networks

- Expressing opinion an important element
  1. Comments (on photographs, status msgs.)  
*'Pritesh Patel loved the pasta he had at Pizza hut today'*
  2. Status messages / tweets  
*'Nokia E51. Become a fan'.*
  3. 'Become a fan' on facebook  
*'4 of your friends are a fan of Ganpati. Become a fan'.*

# What? Comments

- In what form does opinion exist on the web?
- Comments everywhere



The screenshot shows a web browser window with the URL <http://timesofindia.indiatimes.com/news/world/us/Obama-pushes-agenda-to-end-nuclear-weapons/articleshow/5047971.cms>. The browser's address bar and tabs are visible. Below the browser window, there is a "WRITE YOUR COMMENT" form. The form includes fields for "Name:", "Location:", "Email Address:", and "Comment:". The "Email Address:" field has a note: "You will receive an email alert as soon as your comment is live". At the bottom of the form, there are "Preview" and "Post comment" buttons. Below the form, there is a navigation bar with icons for "Print", "Email", "Discuss", "Bookmark/Share", "Save", "Comment", and "Text Size: A | A".

From: [www.timesofindia.com](http://www.timesofindia.com)

# What? Comments

- Two types of comments:
  - Comments about the article/ blogpost:
    - *Very well-written indeed...*
  - Comments about the topic of the article:
    - *I agree with you.. I used to love \*\*'s movies at a point of time but these days all he comes out with is trash. <Often leads to a conversation>*
- ( - Comments about the blogger:
  - *If you think Shahid Kapoor is ugly, go buy glasses. While you are at it, buy yourself a brain too*
- )

# Outline

## Lecture 1

### Motivation & Introduction

Challenges of SA: *Why SA is non-trivial*

Variants of SA: *What forms does it exist in?*

Opinion on the web: *Is doing SA really worth it?*

### Classifiers for SA

Fundamentals of supervised approaches

Standard ML techniques

Comparing different classifiers for SA

## Lecture 2

### Approaches to SA

### Applications

# What is classification?

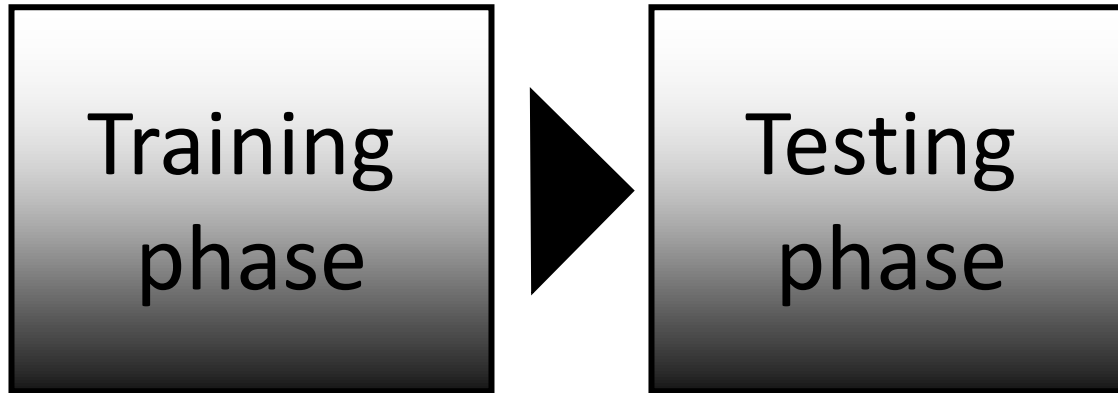
A machine learning task that deals with identifying the class to which an instance belongs

A classifier performs classification





# Classification learning



Learning the classifier  
from the available data  
'Training set'  
(Labeled)

Testing how well the classifier  
performs  
'Testing set'

# Testing phase

## Methods:

- Holdout ( $2/3^{\text{rd}}$  training,  $1/3^{\text{rd}}$  testing)
- Cross validation (n – fold)
  - Divide into n parts
  - Train on (n-1), test on last
  - Repeat for different permutations
- Bootstrapping
  - Select random samples to form the training set

# Outline

## Lecture 1

### Motivation & Introduction

Challenges of SA: *Why SA is non-trivial*

Variants of SA: *What forms does it exist in?*

Opinion on the web: *Is doing SA really worth it?*

### Classifiers for SA

Fundamentals of supervised approaches

Standard ML techniques

Comparing different classifiers for SA

## Lecture 2

### Approaches to SA

### Applications

# ML-based classifiers

- Naïve Bayes
- Maximum Entropy
- SVM
- Committee-based classifiers

# Naïve Bayes classifiers

- Based on Bayes rule
- Naïve Bayes : Conditional independence assumption

$$P(C_j | X) = \frac{P(X | C_j) \cdot P(C_j)}{P(X)}$$

$$P(X | C_j) = \prod_{k=1}^d P(x_k | C_j)$$

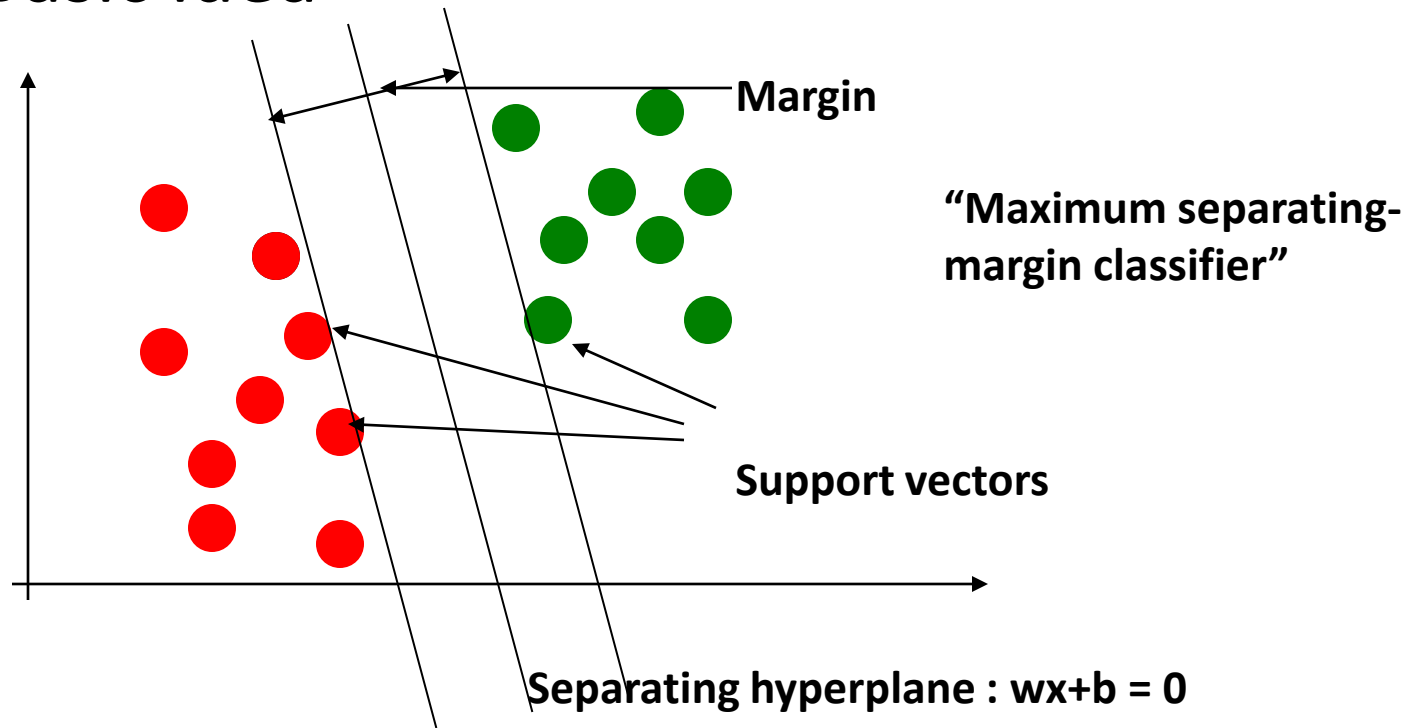
# Maximum Entropy

$$P_{\text{ME}}(c \mid d) := \frac{1}{Z(d)} \exp \left( \sum_i \lambda_{i,c} F_{i,c}(d, c) \right)$$

$$F_{i,c}(d, c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases} .$$

# Support vector machines

- Basic idea



# Multi-class SVM

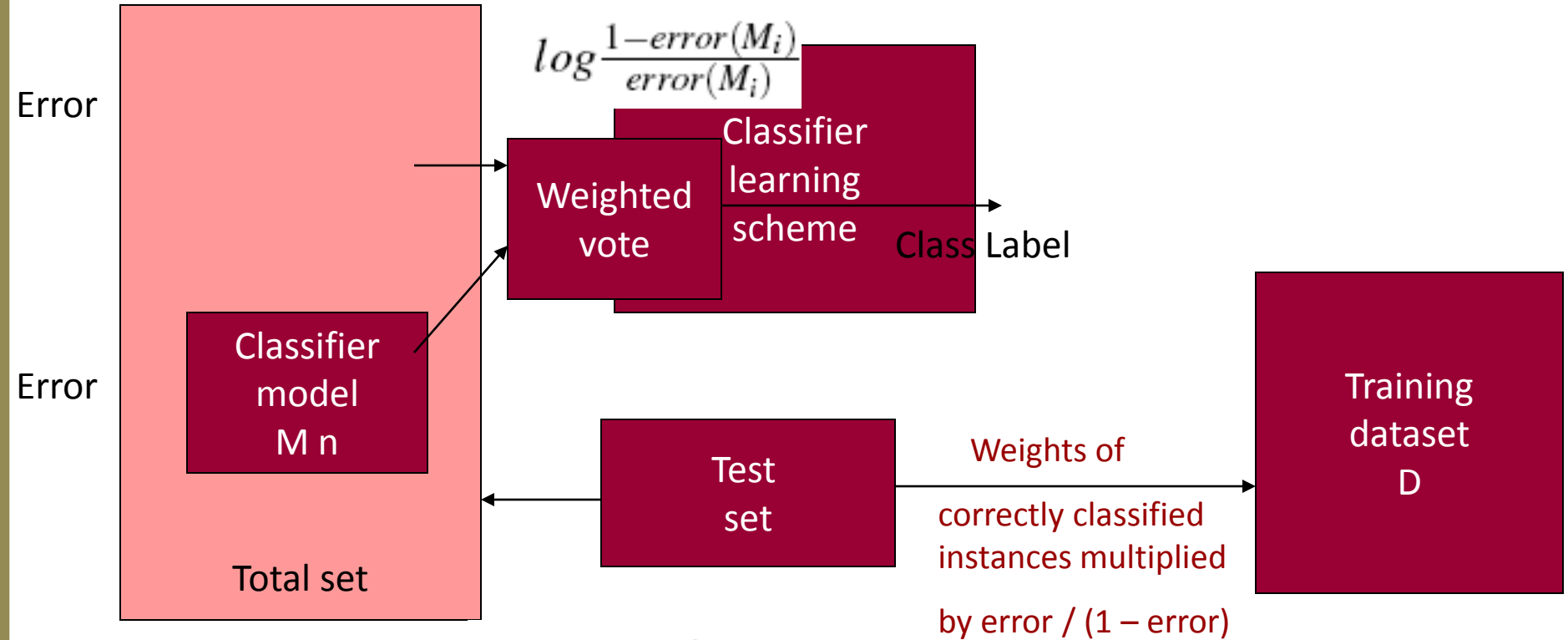
- Multiple SVMs are trained:
  - True/false classifiers for each of the class labels
  - Pair-wise classifiers for the class labels



# Combining Classifiers

- 'Ensemble' learning
- Use a combination of models for prediction
  - Bagging : Majority votes
  - Boosting : Attention to the 'weak' instances
- Goal : An improved combined model

# Boosting (AdaBoost)



San  $error(M_i) = \sum_j^d w_j \times err(X_j)$  by the bootstrap

# Outline

## Lecture 1

### Motivation & Introduction

Challenges of SA: *Why SA is non-trivial*

Variants of SA: *What forms does it exist in?*

Opinion on the web: *Is doing SA really worth it?*

### Classifiers for SA

Fundamentals of supervised approaches

Standard ML techniques

Comparing different classifiers for SA

## Lecture 2

### Approaches to SA

### Applications

# Task Definition

- Marking reviews as positive or negative on the document level
- List-based classifiers
- ML-based classifiers
  - Term presence/Term frequency
  - Unigram/bigram
  - Adjectives

# Results

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	<b>78.7</b>	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	<b>82.9</b>
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	<b>82.7</b>
(4)	bigrams	16165	pres.	77.3	<b>77.4</b>	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	<b>81.9</b>
(6)	adjectives	2633	pres.	77.0	<b>77.7</b>	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	<b>81.4</b>
(8)	unigrams+position	22430	pres.	81.0	80.1	<b>81.6</b>

Compared to list-based classifiers (58-69%)

# Analysis

- On the surface level, ML-based classifiers do better than lexical-based classifiers
  - Worse than a human being
- Discourse understanding important to tackle thwarted expressions

# Outline

## Lecture 1

### Motivation & Introduction

Challenges of SA: *Why SA is non-trivial*

Variants of SA: *What forms does it exist in?*

Opinion on the web: *Is doing SA really worth it?*

### Classifiers for SA

Fundamentals of supervised approaches

Standard ML techniques

Comparing different classifiers for SA

## Lecture 2

### Approaches to SA

### Applications