Sentiment Analysis and Opinion Mining

Chenghua Lin
Computing Science
University of Aberdeen

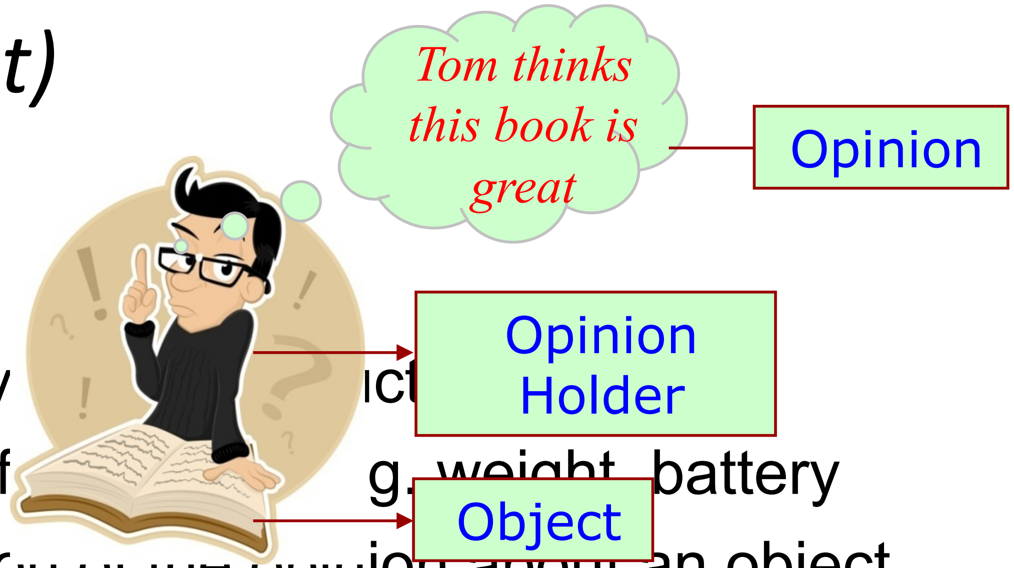
What is Sentiment Analysis

- Sentiment
  - A thought, view, or attitude, especially one based mainly on emotion instead of reason
- Sentiment Analysis
  - Computational treatments for discovering opinions and attitudes expressed in text by opinion holders (e.g. positive vs. negative)
  - Can be generalised to richer emotion dimensions (e.g., Joy, Sadness, Fear, Anger, etc.)

Definitions

- Modelling opinion as a quintuple: (Bing Liu 2012)

$$(o, a, so, h, t)$$



Where

- ***o*** is an object or target entity … ict
- ***a*** is the aspect or attribute of … g. weight, battery
- ***so*** is the sentiment orientation of the opinion about an object or aspect
- ***h*** is an opinion holder
- ***t*** is the timestamp when an opinion is expressed.

Different Types of Opinions

- Explicit opinion: opinion or sentiment directly expressed on a target object
  - "*My mood is really bad.*"
  - "*His basketball skill is really amazing*!"
- Implicit opinion: objective text implying opinion or attitude
  - "*The new bike fell apart within two days.*"
  - "*He has learnt a lot from this course*."
- Sarcasm: a sharp or bitter remark, usually conveys the opposite of their literal meaning (context-dependent)
  - **Context**: "*The food is totally burned*! (very angry)"
  - "*You really did a great job!*"

Why Sentiment Analysis

- Web 2.0 and the social web
  - has facilitated rich user-generated contents
  - contains valuable information for both business and end-users
- We have a *decision support* need and an *operational* need
  - ➢ **Marketers and governments**
  - Finding out customers' opinions about their products/services
  - Tracking how these opinions evolve over time
  - Accessing public opinion polls on political campaigns
  - ➢ **Consumers**
  - Decision support for purchasing and recommendation
  - What are people saying about X versus Y

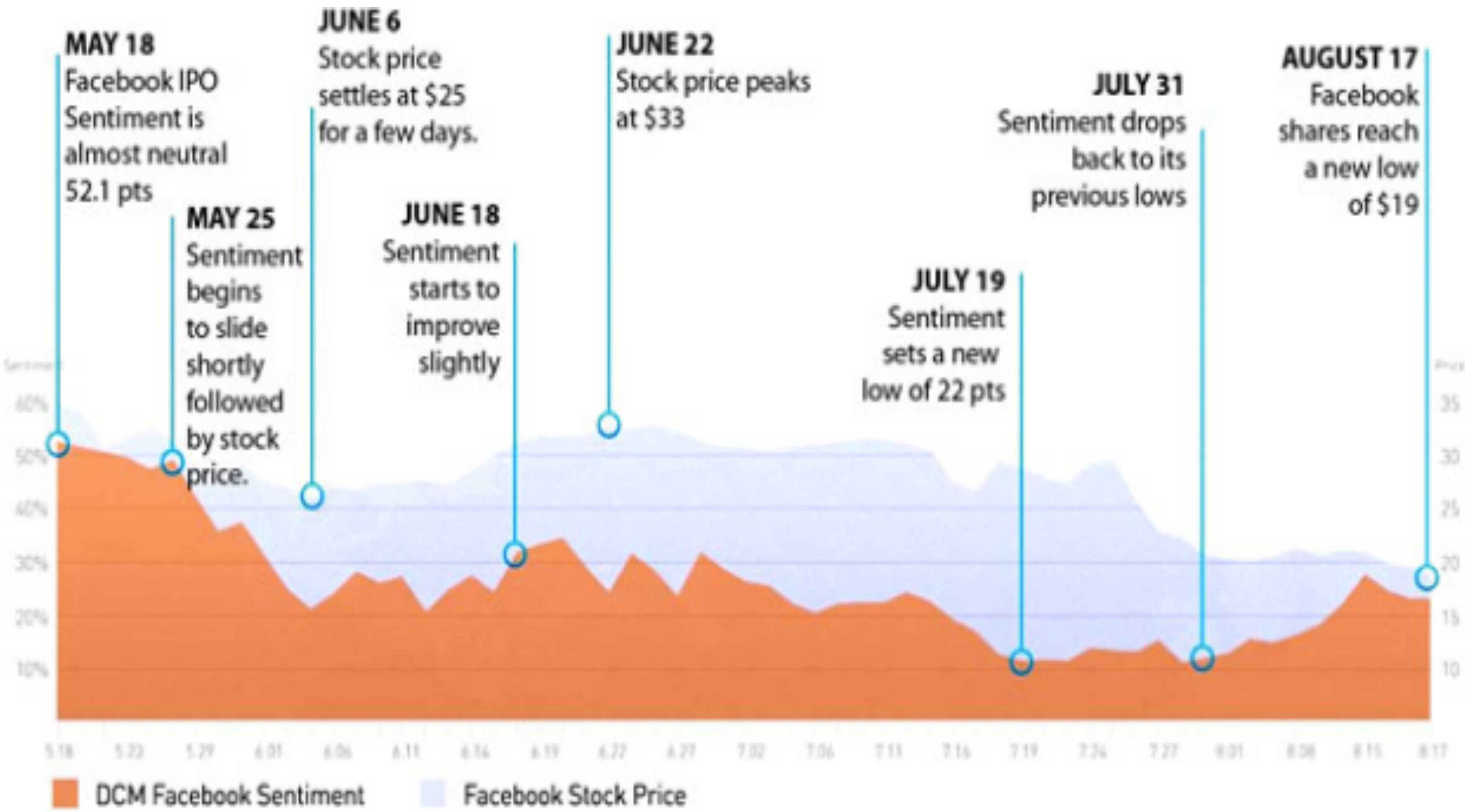
Some Practical Examples

Product Review Insights

- What are people's opinions about this product?
- What are the pros and cons?

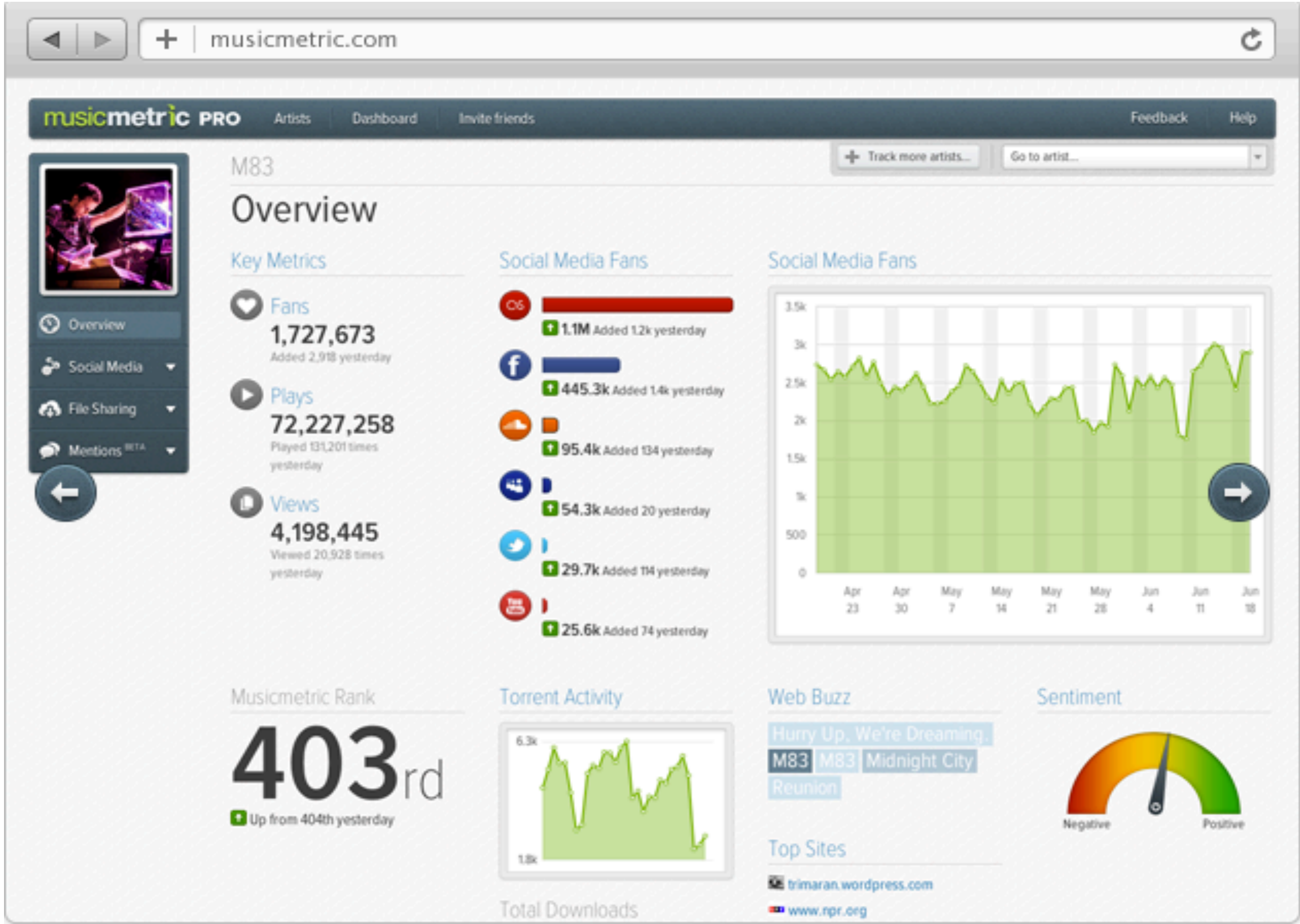
Financial Marketing

- Use general Facebook sentiment to predict the company's stock performance. (http://www.dailyfinance.com/)



Brand and Consumer Perception

- Music artists analytics: provide aggregated sentiment statistics for artists, songs or albums over all reviews collected online.



Sentiment Analysis Tasks

The Big Picture

Sentiment classification	Is a document/sentence positive or negative?
Subjectivity detection	Whether given text expresses opinions (subjective) or reports facts (objective)
Opinion holder/target identification	Who express a specific opinion? What features of the iPhone 5 do customers like?
Opinion summarisation	Summarise opinions over multiple review documents towards a certain product
Opinion retrieval	How do people think of iPhone5?
Sentiment dynamics prediction and tracking	How does people’s views on Mac change over time?
Opinion spam detection	Opinion spam detection: Identify fake/untruthful reviews.

Sentiment Classification

Sentiment Classification

- **Goal:** classify the overall sentiment orientation expressed in a given text, i.e. positive, negative or (possibly) natural.
- Classification can involve different levels of granularity
  - Document-level
  - Sentence-level
  - Word/phrases-level
- Traditional sentiment classification techniques
  - Lexicon-based approaches
  - Corpus-based approaches

Sentiment Classification Techniques

- Lexicon-based approaches
  - Use sentiment lexicons as prior knowledge
  - Unsupervised/weakly supervised learning

- Corpus-based approaches
  - Annotated corpus with class labels available (e.g. positive or negative)
  - Supervised learning, e.g., Naïve Bayes (NB), SVMs, Maximum Entropy Model (MaxEnt), etc.

Preprocessing

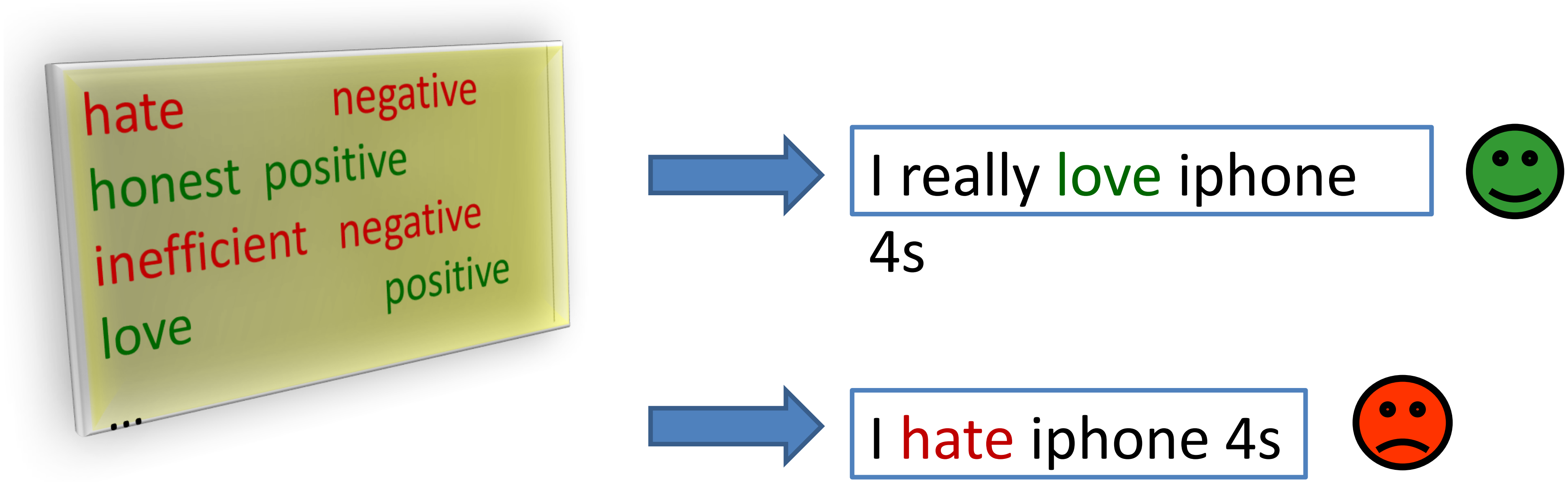
- Text preprocessing
  - An essential part of any NLP system
  - Segment text into appropriate unit (e.g. characters, **words**, sentences, etc.) and prepare in appropriate forms (e.g. canonical form) before passing for further processing
  - Also called text tokenization and normalization
- Easier in some languages (e.g. English) and highly non-trivial for the languages without space (e.g. Chinese).
- Have a great impact on NLP system performance, e.g. speed, accuracy, etc.

Preprocessing

Typical steps of preprocessing in English:

- **Cleaning**, e.g. HTML tags. But sometimes can be helpful, e.g. tags preserving document structures like headings, sections.
- **Text unit segmentation**: normally white space and punctuations as word boundaries. Instances like '*e-mail*', '*aren't*', can be problematic.
- **Stop word removal**: (1) remove the most frequent words that do not carry much meaning. E.g., '*the'*, '*and'*, '*a'*, etc. (2) can greatly reduce corpus size.
- **Stemming**: convert the inflected words to their stem, e.g., '*stocking*', '*stocks*', '*stocked*' → '*stock*'.
  - Porter stemmer
  - Reduce vocabulary size
  - May collapse words with different meaning into the same stem – '*pass*', '*passe*' → '*pass*'

Lexicon-Based Approaches



Sentiment lexicon

- Use sentiment words as reference features for polarity detection
- Does not rely on labelled data for training

Example: Unsupervised Learning

Unsupervised sentiment classification (Turney 2002)

- Goal: polarity classification on reviews, i.e. positive or negative
- Data: reviews from epinions.com: automobiles, banks, movies, and travel destinations.
- Two polarity words: '*excellent*', '*poor*'
- Approach: point-wise mutual information (PMI)

Step 1

Step 1

- Part-of-speech (POS) tagging
- Apply POS patterns -- extract two consecutive words (two-word phrases) from reviews if their tags conform to some given patterns, e.g., (1) JJ, (2) NN.

	First word	Second word	Third word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Step 2

Step 2: Estimate the sentiment orientation
phrases

measures the degree of
statistical dependence
between words

- Use Pointwise mutual information

$$\mathrm{PMI}(w_1, w_2) = \log_2\left(\frac{p(w_1 \wedge w_2)}{p(w_1)p(w_2)}\right)$$

- Sentiment orientation (SO):

$$\mathrm{SO}(\mathit{phrase}) = \mathrm{PMI}(\mathit{phrase}, \text{“excellent”}) - \mathrm{PMI}(\mathit{phrase}, \text{“poor”})$$

- In practice, approximately compute PMI and SO using public search engine AltaVista.

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

$$\mathrm{SO}(\mathit{phrase}) = \log_2\left(\frac{\text{hits}(\mathit{phrase}\ \ \text{NEAR}\ \ \text{“excellent”})\ \text{hits}(\text{“poor”})}{\text{hits}(\mathit{phrase}\ \ \text{NEAR}\ \ \text{“poor”})\ \text{hits}(\text{“excellent”})}\right)$$

Step 3

Step 3:

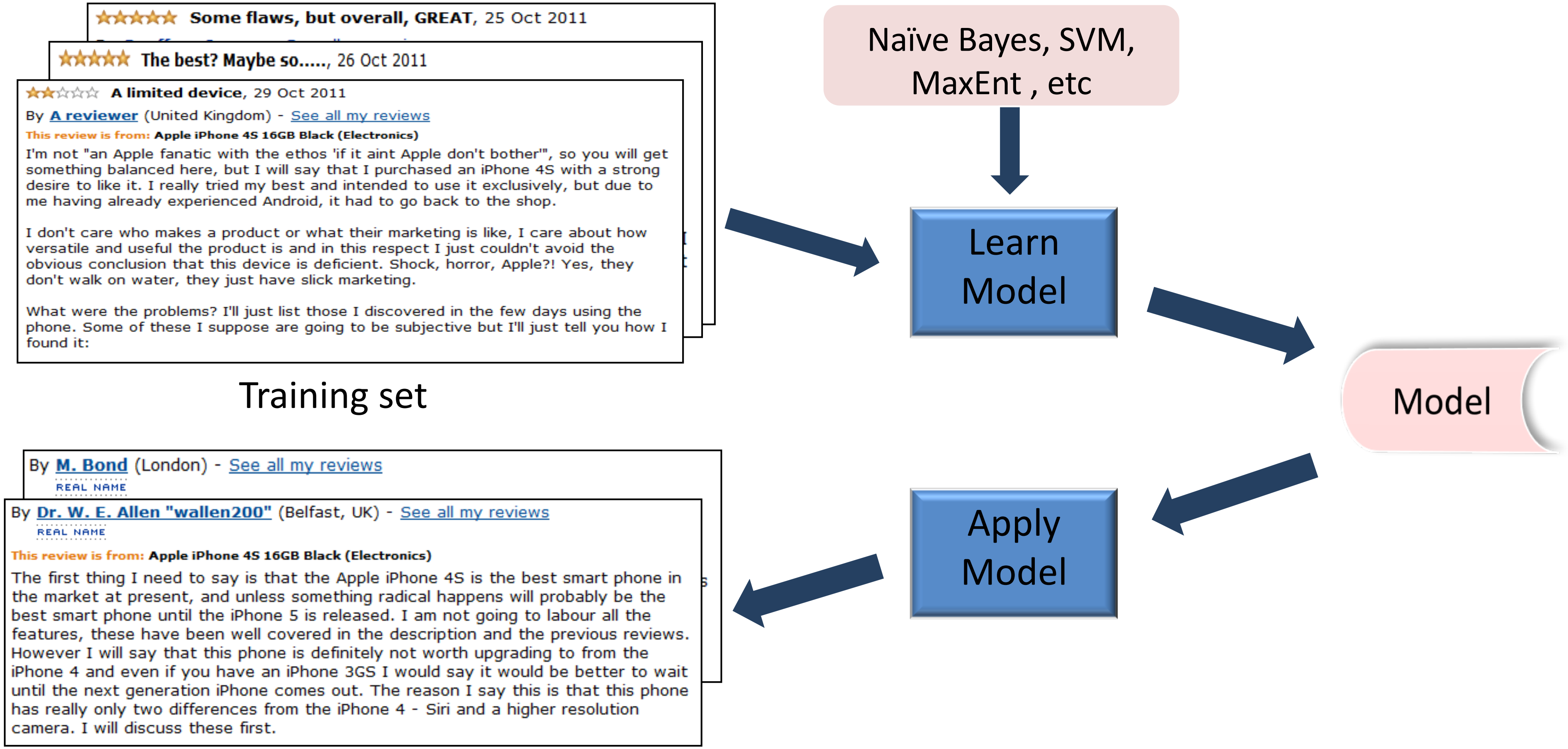
- Compute the average SO of all phrases
- Classify the review as positive if average SO is positive, negative otherwise.

- Final classification accuracy:
  - Automobiles: 84%
  - Banks: 80%
  - Movies: 65.83
  - Travel destinations: 70.53%

Corpus-based Approaches

- **Basic idea**: treat sentiment classification as a binary classification problem with two topics, i.e. 'positive' and 'negative'.
- Supervised classification algorithms
  – Naïve Bayes (NB)
  – Support Vector Machines (SVM)
  – Maximum Entropy (MaxEnt), etc.
- Commonly used benchmark dataset
  – Movie reviews (http://www.cs.cornell.edu/people/pabo/movie-review-data/)
  – Product reviews (http://www.cs.jhu.edu/~mdredze/datasets/sentiment/)

Corpus-based methods (cont.)



Test set

- Rely on syntactic or co-occurrence patterns in large text corpora

Example: Supervised Learning

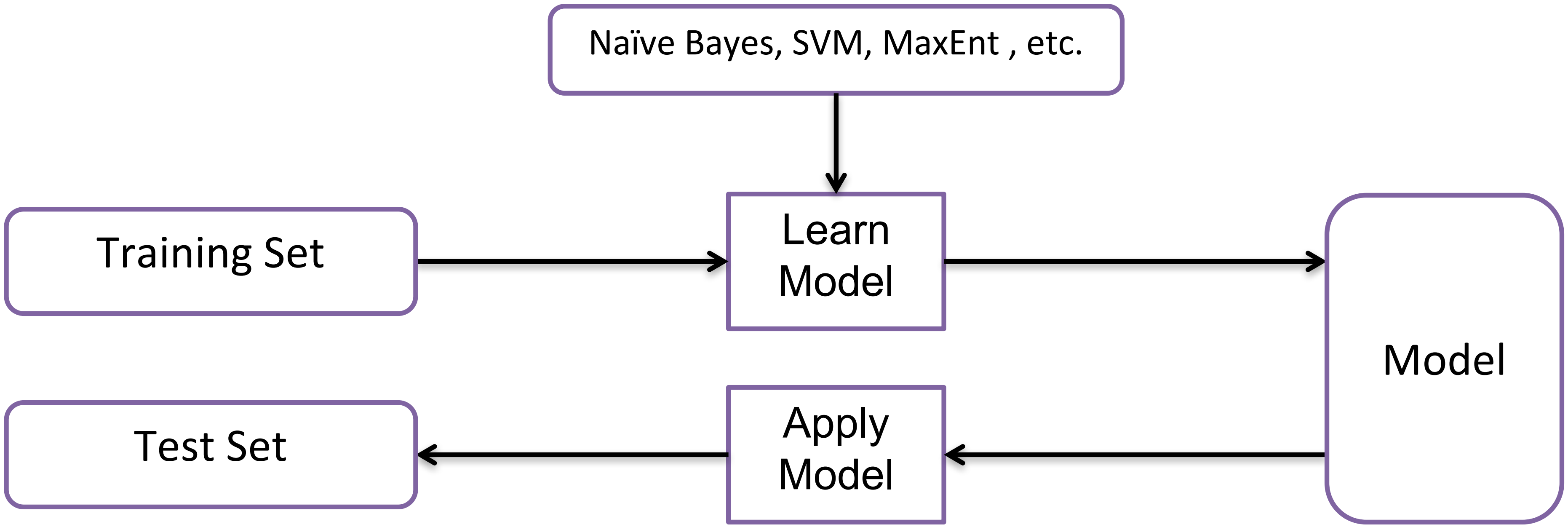
Supervised machine learning algorithms for sentiment classification (Pang et al, 2002, 2004)

- **Goal**: predict the sentiment orientation of a document as positive or negatives
- **Algorithms**
  - Navie Bayes (NB)
  - Support Vector Machines (SVMs)
  - Maximum entropy (MaxEnt)
- **Data and setting**
  - 700 positive (4-5 stars) and 700 negative (1-2 stars) reviews
  - 3-fold cross-validation
  - No stemming or stopword removal

Other Sentiment Classification Subtasks

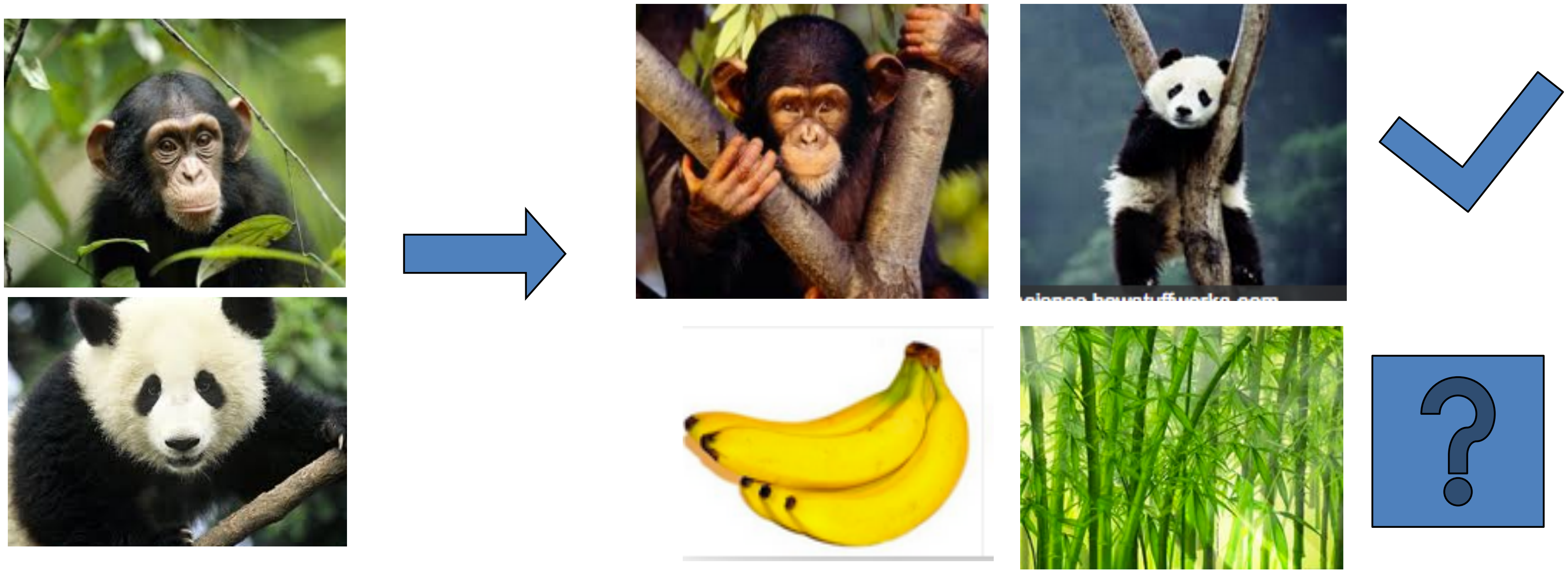
- Domain adaptation
- Contrastive opinion mining
- Multi-level sentiment classification
- Sentiment dynamics

Traditional Classification



An Important Assumption

- Training and unseen (test) data
  - in the same feature space
  - follow the same distribution



Training data

Unseen data

In the Real World …

- More often than **NOT**, the aforementioned assumptions do not hold
- May not have sufficient training data in the domain of interest (target)
- Rather, we have a lot of labelled data available in other domains (source)
- Source domain data exhibit different distributions or feature space to the target domain
- Labelling cost is expensive

Domain adaptation

- May not have sufficient training data in the domain of interest (target)
- Rather, we have a lot of labelled data available in other domains (source)
- Source domain data exhibit different distributions or feature space to the target domain
- Labelling cost is expensive

Some Early Attempts

- Aue and Gamon(2005): “*Explore customizing sentiment classifiers to new domains*”
- Data
  - Movie review data (**movie**)
    - 1000 positive and 1000 negative reviews from movie databases
  - Book review data (**book**)
    - 1000 positive and 1000 negative book reviews from the web
  - Product Support Services web survey data (**pss**)
    - 2564 examples of positive feedback and 2371 examples of negative feed- back
  - Knowledge Base web survey data (**kb**)
    - 6035 examples of “bad” feedback and 6285 examples of “good” feedback.

The Cross-Domain Issues

(Aue and Gamon 2005)

- Sentiment classifier trained on one domain do not generalize well to other domains

	movie	book	kb	pss
movie	**90.45**	70.29	57.59	61.36
book	72.08	**79.42**	59.28	66.59
kb	57.1	58.62	**77.34**	81.42
pss	52.16	55.33	70.48	**83.73**

Table 1: Best results of SVM classifiers within and across domains.

kb: Knowledge base web survey data

pss: product support services web survey data

Cross-Domain Classification

- Strategies
  1. Mixing available labelled data from all source domains → trained a single classifier **(data fusion)**
  2. Similar to (1), but limiting the feature set to those observed in the target domains **(data fusion + feature engineering)**
  3. Train a single classifier for each domain using labelled data → ensemble classifiers **(classifier fusion)**
  4. NaiveBayes-EM: trained on **in-domain labelled** data (small amounts) + **in-domain unlabelled** data (large amounts)

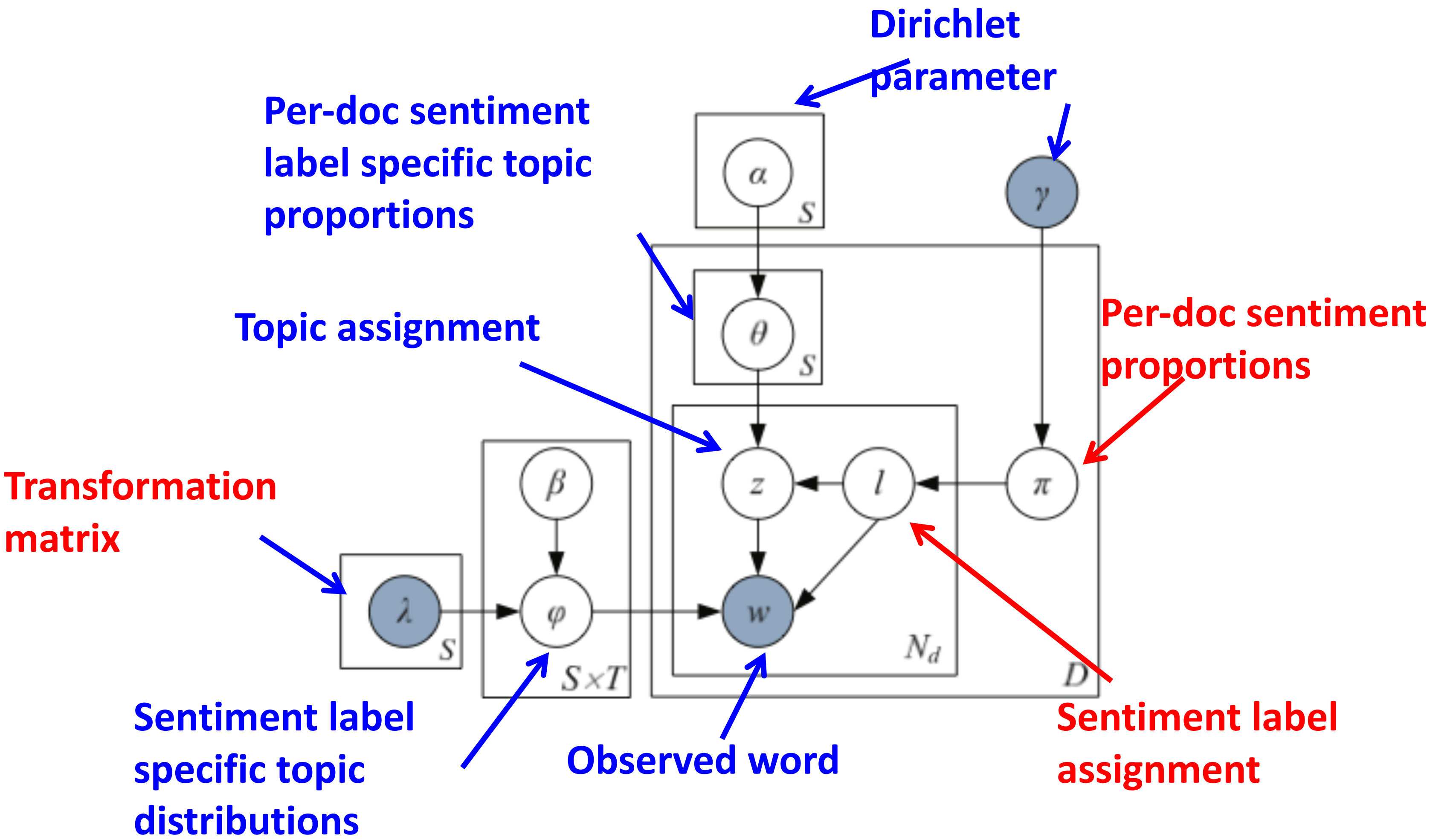
Topic Model based Approaches

Joint Sentiment-Topic (JST) Model

(Lin & He 2009, Lin et al 2011)

- **JST**: A generative model by extending latent Dirichlet allocation (LDA) with an additional sentiment layer
  - Detect document-level sentiment and extract sentiment-bearing topics simultaneously from text
  - Weakly-supervised learning without using labelled data, i.e. only a small set of domain independent sentiment lexicon were used as polarity features
  - The automatically extracted sentiment-bearing topics can be used as succinct summaries for large amount of reviews

Joint Sentiment-Topic (JST) Model



- Detect the overall sentiment for a review/document.
- Automatically extract sentiment-bearing topics as succinct summaries for data

Dataset

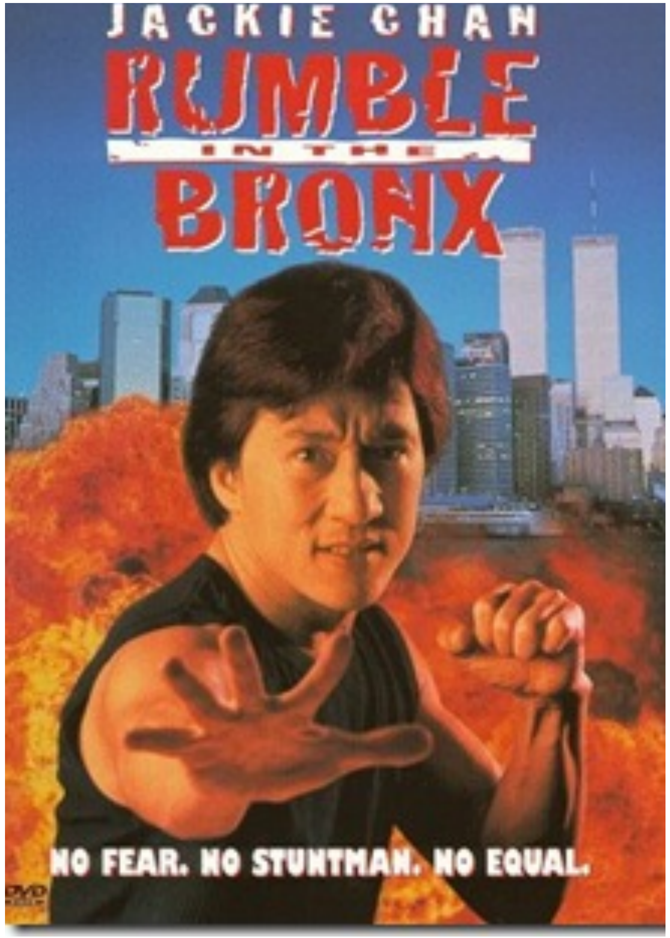
- Movie review (MR) dataset version 2.0[1]
  - 1000 positive and 1000 negative movie reviews from IMDB
  - 30 sentences per document on average
- Subjectivity movie review dataset
  - Based on the MR dataset
  - Strip the sentences do not contain subjective information
- Multi-domain sentiment dataset
  - 4 types of reviews from Amazon: Book, DVD, Electronics and Kitchen
  - 1000 positive and 1000 negative examples each domain

[1] http://www.cs.cornell.edu/People/pabo/movie-review-data/

Topic Extraction (DVD domain)

Jackie Chan Movie

action
good
fight
right
scene
chase
hit
art
martial
stunt
chan
brilliant
hero
style
chines



Topic Extraction (Electronic domain)

Flash disc damage

drive
fail
data
complet
lose
failur
recogn
backup
poorli
error
storag
gb
flash
disast
yesterdai

Applying JST to Review Data

Customer Reviews

Amazon Kindle Keyboard Leather Cover, Black

855 Reviews

5 star:	(594)
4 star:	(167)
3 star:	(47)
2 star:	(22)
1 star:	(25)

Average Customer Review

(855 customer reviews)

Share your thoughts with other customers

Create your own review

46 of 48 people found the following review helpful:

Nice quality but......, 4 Sep 2010

By **C. Knight** (UK) - See all my reviews

REAL NAME

This review is from: **Kindle Leather Cover, Black (Fits 6" Display, Latest Generation Kindle) (Accessory)**

Not a bad cover. Nice leather, neat design , but overpriced, particularly when you consider the kindle is only 3 times the prices of this cover. And wouldnt it have been nice to make it left or right handed. The kindle is great having the page switches on both sides. but the cover makes the left hand switches harder to use. I do think the kindle should have come packaged with a cheap sleeve and feel this is a bit of a marketing rip off , (there is no way I can walk round with the kindle unprotected). But, have to say, it does the job. Just well overpriced.

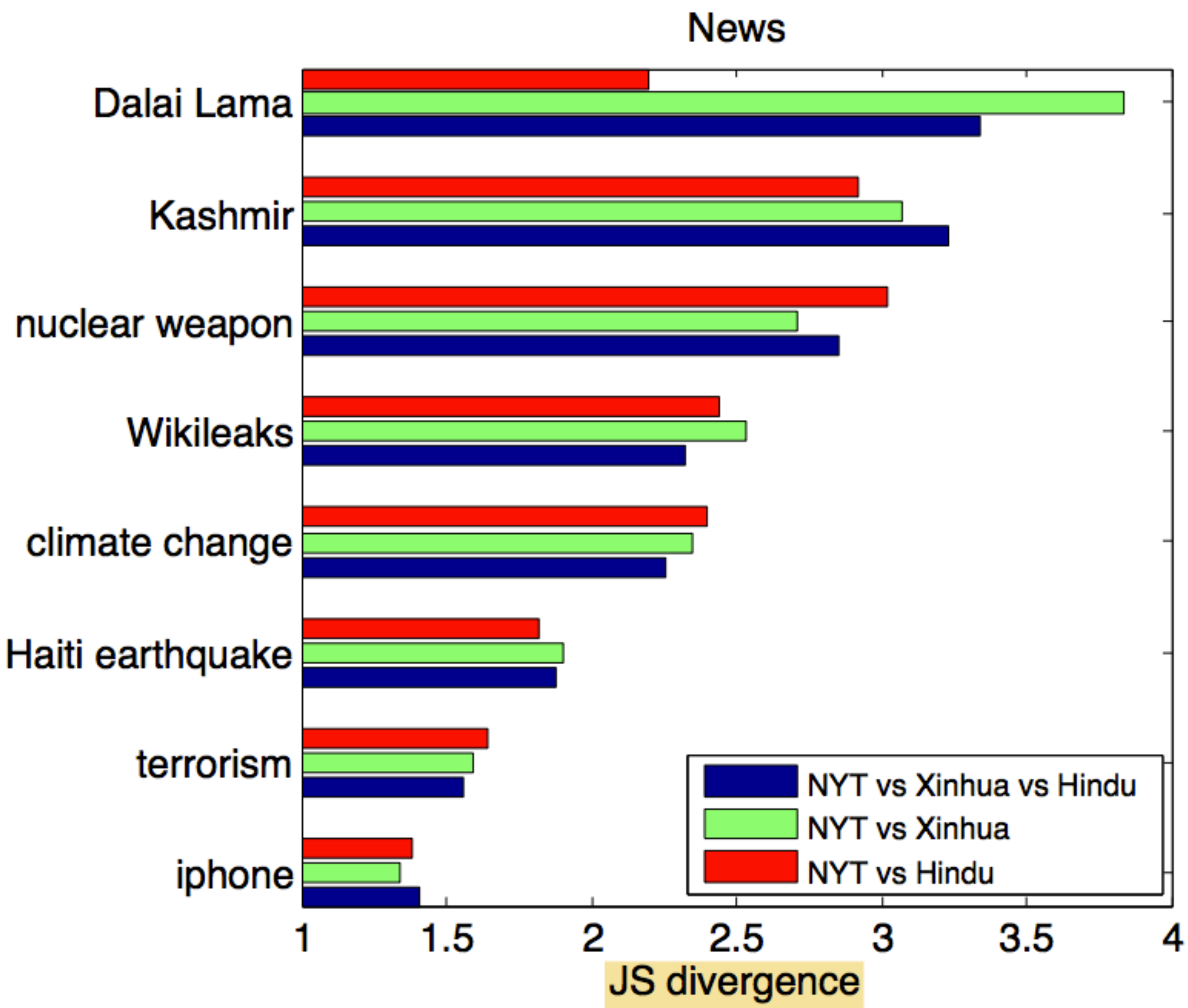
Negative sentiment topic

price
cover
high
expensive
kindle
overpriced
...

Contrastive Opinion Mining

		New York Times		Xinhua News		The Hindu	
TOPIC 40							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
peace	0.0573	dissident	0.0116	convicted	0.0163	jailed	0.0168
prize	0.0533	awarded	0.0101	arrogant	0.0114	dissident	0.0077
nobel	0.0425	democratic	0.0078	political	0.0092	criticised	0.0062
liu	0.0391	imprisoned	0.0068	interfere	0.0092	unaware	0.0062
committee	0.0281	pro-democracy	0.0049	internal	0.0085	imprisoned	0.0046
TOPIC 54							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
iran	0.2209	military	0.0765	peaceful	0.1065	diplomatic	0.0713
program	0.0391	impose	0.0623	diplomatic	0.0128	international	0.0496
tehran	0.0334	stop	0.0442	negotiate	0.0121	military	0.0367
uranium	0.0305	diplomatic	0.0307	civilian	0.0113	constructive	0.0299
ahmadinejad	0.0195	financial	0.0225	unilateral	0.0109	regional	0.0186
TOPIC 68							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
kashmir	0.0404	ethnic	0.0147	indian-controlled	0.0205	democratic	0.0275
pakistan	0.0388	killed	0.0147	moderate	0.0198	civil	0.0215
constitution	0.0222	disputed	0.0145	end	0.0162	insurgent	0.0201
violence	0.0193	peaceful	0.0145	infiltrate	0.0156	separate	0.0195
valley	0.0157	tibetan	0.0142	bilateral	0.0151	military	0.0184
TOPIC 81							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
china	0.2414	economic	0.1133	economic	0.1175	economic	0.1147
chinese	0.1063	state-run	0.0165	western	0.0520	communist	0.0193
beijing	0.0672	rising	0.0159	peaceful	0.0180	growing	0.0186
government	0.0213	manipulate	0.0156	positive	0.0171	rising	0.0180
currency	0.0109	controlled	0.0154	developing	0.0160	territorial	0.0179

Contrastive Opinion Mining



Corpus for Sentiment Analysis

Loads of training data available:

- Movie review (http://www.cs.cornell.edu/home/llee/data/)
- Multi-domain Amazon product review (http://www.cs.jhu.edu/~mdredze/datasets/sentiment/)
- Multiple-aspect restaurant reviews (http://people.csail.mit.edu/bsnyder/naacl07)
- The Stanford Twitter Corpus (http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip)
- The SemEval-2013: Sentiment Analysis in Twitter evaluation campaign (or competition) dataset (http://www.cs.york.ac.uk/semeval-2013/task2/index.php?id=data)
- … …

Sentiment Lexicons

- MPQA Subjectivity Cues (http://mpqa.cs.pitt.edu/)
  - 2718 positive, 4912 negative words
- Bing Liu's Opinion Lexicon (http://www.cs.uic.edu/~liub)
  - 2006 positive and 4783 negative words
- SentiWordNet (http://sentiwordnet.isti.cnr.it/)
  - a polarity dictionary with strength scores
  - Automatically annotate WordNet synsets for degrees of sentiment intensity
- The General Inquirer (http://www.wjh.harvard.edu/~inquirer/homecat.htm)
  - 1915 positive and 2291 negative words

Software Packages/Libraries

- Stanford NLP software (http://nlp.stanford.edu/software/)
  - Statistical NLP toolkits for various major computational linguistics problems
- Mallet (http://mallet.cs.umass.edu/)
  - A Java-based package for statistical natural language processing
- LingPipe (http://alias-i.com/lingpipe/)
  - A Java tool kit for processing text using computational linguistics
- OpenNLP （http://opennlp.apache.org/）
  - The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text.
- GATE (http://gate.ac.uk/)
  - Open source software for text analysis
- Python NLTK (http://nltk.org/)
  - A Python platform for processing human language data
- RTextTools (http://www.rtexttools.com/)
  - A machine learning library for text classification in **R**