Sentiment Identification

Lecture 9: Text Analytics

Mark Keane, Insight/CSI, UCD

Selling Things

stock-markets

social media

science

news

polls

***

sentiment-id

sentiment-use

time-series

summaries

***

VSMs

Classifiers

Clustering

cosine

jaccard

dice

levenschtein

TF-IDF

LLR

PMI

Entropy

simple frequencies

pre-processed text items of some sort…

Basically It's…

- Words express various sentiments, feelings, opinions, biases, reactions…

- Tracking sentiment words tells us what people feel about these things, their opinions…

- Opinion mining, sentiment analysis, subjectivity analysis, appraisal extraction…

Overview

- Identifying Sentiment Terms (Lect-9)
- Using Sentiments to Do Things (Lect-10)
- Key Examples (both)
- Some Implementations (both)

What is an Opinion?

- An opinion is simply a positive or negative sentiment, view, attitude, emotion, or appraisal of an entity or an aspect of the entity, from an opinion holder (Hu and Liu 2004; Kim and Hovy 2004; Wiebe et al 2005)

"I love iPhones"; "Bill says Windows is crap"

"The screens on iPhones are terrible"; "the Launcher feature in Windows is great"

Opinions Types

- ***Regular***: opinion about a target entity
  - Direct: “I hate iPhones”
  - Indirect: “IKEA chairs changed my life”
- ***Comparative***: comparing two entities
  - Direct: “iPhones are way better than Nokia”
  - Indirect: “Moving from PCs to Macs, changed my life…”

Bing Liu. (2010) “Sentiment Analysis and Subjectivity." *Handbook of Natural Language Processing*.

What's a Sentiment?

- Is what is expressed in an opinion: positive, negative or neutral (I have no opinion)
- Polarity / valency of a sentiment is its positive or negative feeling tone
- Also called opinion orientation, semantic orientation, sentiment polarity

More Formally…

An *opinion* is a quintuple

$$(e_j, a_{jk}, so_{ijkl}, h_i, t_l),$$

where

- e_j is a target entity.
- a_{jk} is an aspect/feature of the entity e_j.
- so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or more granular ratings.
- h_i is an opinion holder.
- t_l is the time when the opinion is expressed.

Bing Liu. (2010) "Sentiment Anlaysis and Subjectivity." *Handbook of Natural Language Processing*.

More Terminology…

An *opinion* is a quintuple

$(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

where

- e_j is a target entity.
- a_{jk} is an aspect/feature of the entity e_j.
- so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or more granular ratings.
- h_i is an opinion holder.
- t_l is the time when the opinion is expressed.

object

topic

attribute, facet

opinion source

Bing Liu. (2010) "Sentiment Anlaysis and Subjectivity." *Handbook of Natural Language Processing*.

The Problem…

- Identifying these quintuples in a text
- Core task is identifying the words that express an opinion and their sentiment
- But, often, not all components of quintuple can be found

Ways to Identify Sentiments

- ***Human Ratings***: of sentiment terms / phrases / documents

- ***Sentiment Lexicons:*** Matching text bits to sentiment word lists and lexicons (built from human judgments)

- ***Sentiment Classifiers:*** Classifying texts to find sentiment features (built from human ratings)

Identifying Sentiment Terms:

NLP Problem

What's the problem

- You could say it's identifying these quintuples in the text, but it is really the problem of NLP

- [NB. we are ignoring more simpler cases of given ratings (thumbs up/down or star ratings as they are not text, *per se*]

What's the problem

- NLP problem is really hard; it really amounts to full NLP, which explains why its avoided

"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"

What's the problem

entity detection or extraction

aspects of entities…

"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"

What's the problem

Opinion detection

Sentiment Detection and Polarity of Sentiment…

"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"

What's the problem

Opinion Holder

"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"

What's the problem

- Diff levels and local/global sentiment:
  - entity or aspect of entity
  - phrase or sentence
  - document (overall, is the review positive?)
- Also note, words can be very subtle:

  "stocks fell" (+'ve) but "inflation fell" (-'ve)

  "it's hard to make a bad picture with Canon"

  "iPhones are good but expensive"

Ways to Identify Sentiments

- ***Human Ratings***: of sentiment terms / phrases / documents

- ***Sentiment Lexicons:*** Matching text bits to sentiment word lists and lexicons (built from human judgements)

- ***Sentiment Classifiers:*** Classifying texts to find sentiment features (built from human ratings)

REM

Identifying Sentiment Terms:

Human Ratings

Expert Ratings: 2 or More

- Get experts / crowd-sourced-non-experts to judge texts / phrases / words being positive or negative
- Extrapolate (in some way) from these ratings to some wider set of matching terms
- Rarely used directly but to support ground-truth for other tasks

Expert Rating: G&K(2011)

- Gerow & Keane (2011): financial articles, 17,713 articles (FT, NYT, BBC), 5.4M words
- Extracted Lemma-Object Pairs (LOPs): “ris* inflation”, “plung* stocks”, “fail* company”
- Raters judge +/-/neutral; both agreement

Gerow, A., & Keane, M. T. (2011). Mining the web for the voice of the herd to track stock market bubbles. *IJCAI-11*. AAAI Press.

G&K(2011): Results

- 3,000 LOP judgements = 49,000 “similar”
- Identical cases and others; “fail* company” gets you “failing company”, “failed company”, “fails company” …
- But, 14% positive, 27% negative, 59% neutral or unsure (residual category)

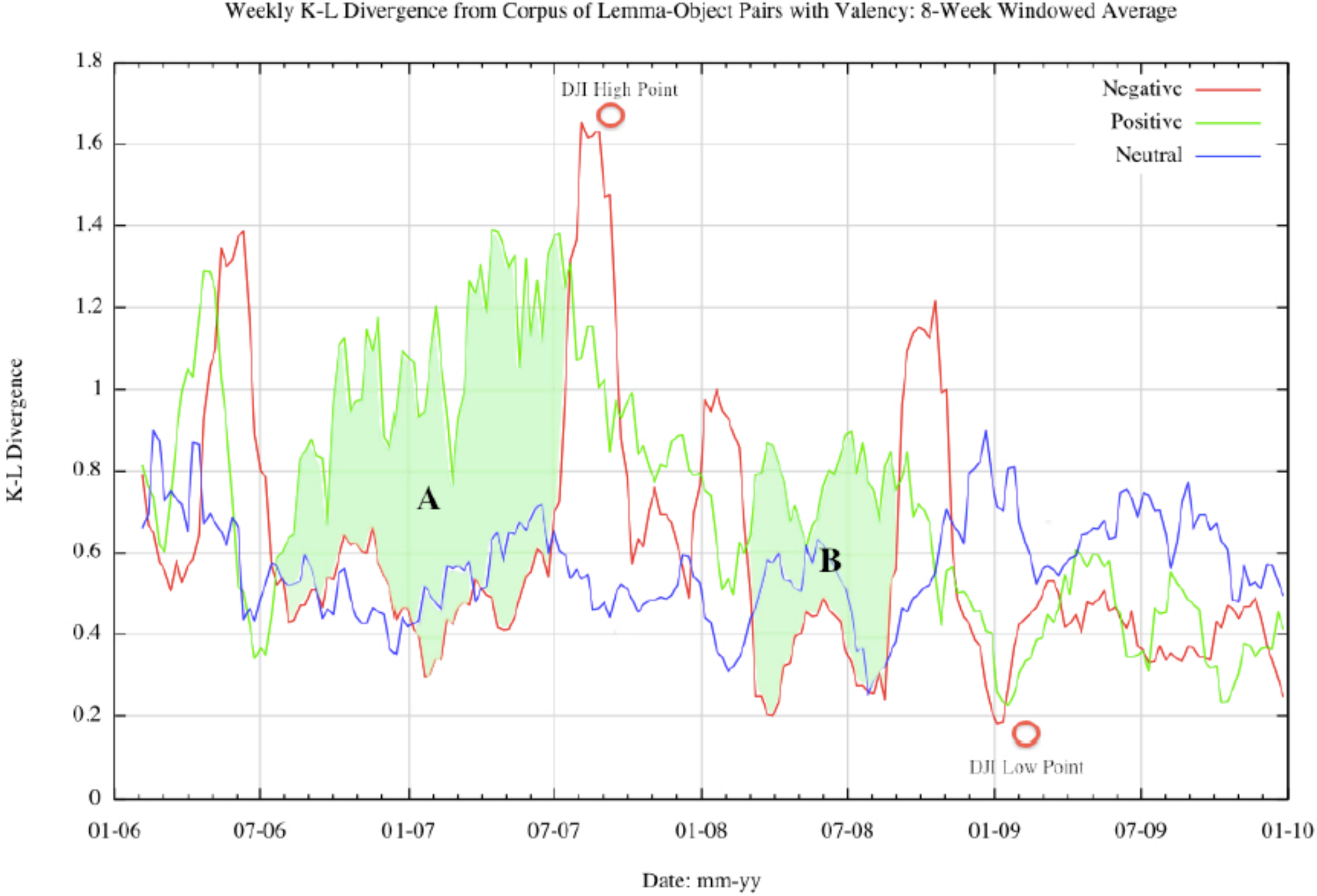


***Figure 2**: Symmetric K-L divergence (8-week windowed mean) of positive, negative, and neutral lemma-object pairs. Note, the two regions, A and B, of distinct positive-negative divergence preceding the 2007 crash and subsequently the beginning of the recovery in 2009.*

G&K(2011): Pros/Cons

- Pro: Sentiment of 1-gram not= 2-gram (rising inflation is **bad**, rising stock prices is *good*)
- Con: Laborious, not feasible for large corpora
- Con: LOP snippet is still too small, mostly neutral
- Con: Disagree ratings are dropped
- Con: Hard on raters

Expert Ratings: Agreement

- G&K used only two experts, often there are more in some tasks

- Raises new issues about how you determine correct answer ?

- Statistics for inter-rater agreement, Pearson Correlation and Cohen's Kappa

Expert Rating: S&M(2007)

- SemEval-2007 Affective Text Task (14)
- 6 Expert raters to annotate news headlines, one six emotions (Anger, Disgust, Fear, Joy, Sadness, Surprise) and valence (+ or -)
- Used expert ratings to determine best system at the Affective-Text-Task

Strapparava, C. and Mihalcea, C (2007). SemEval-2007 Task 14: Affective Text In *Proc. of SemEval-*2007.

Expert Rating: S&M(2007)

- Need to know if rates agree…Methods:
  - Take the majority judgement or >4 agreed
  - Pearson's correlations between raters
  - or Cohen's Kappa or Fleiss' Kappa

Strapparava, C. and Mihalcea, C (2007). SemEval-2007 Task 14: Affective Text In *Proc. of SemEval-*2007.

Agreement: Pearsons

2.3 Inter-Annotator Agreement

We conducted inter-tagger agreement studies for each of the six emotions and for the valence annotations. The agreement evaluations were carried out using the Pearson correlation measure, and are shown in Table 1. To measure the agreement among the six annotators, we first measured the agreement between each annotator and the average of the remaining five annotators, followed by an average over the six resulting agreement figures.

EMOTIONS	
Anger	49.55
Disgust	44.51
Fear	63.81
Joy	59.91
Sadness	68.19
Surprise	36.07
VALENCE	
Valence	78.01

Table 1: Pearson correlation for inter-annotator agreement

PEARSON =PEARSON(array1,array2)

	A	B	C	D	E	F
1	Pearson's Correlation					
2						
3			Rater-1	Rater-2	Rater-3	Mean Rating
4						
5	headline1	Anger	20	25	0	*15.0*
6	headline2	Anger	90	80	0	*56.7*
7	headline3	Anger	0	10	10	*6.7*
8						
9	headline1	Surprise	80	80	0	*53.3*
10	headline2	Surprise	0	10	0	*3.3*
11	headline3	Surprise	0	0	40	*13.3*
12						
13	headline1	Disgust	10	10	90	*36.7*
14	headline2	Disgust	10	10	100	*40.0*
15	headline3	Disgust	100	80	100	*93.3*
16		...				
17						
18						
19	ANGER		SURPRISE		DISGUST	
20	R1 X R2	1.00	R1 X R2	0.99	R1 X R2	1.00
21	R1 X R3	-0.67	R1 X R3	-0.50	R1 X R3	0.50
22	R2 X R3	-0.67	R2 X R3	-0.60	R2 X R3	0.50
23	R1 X MR	1.00	R1 X MR	0.98	R1 X MR	1.00
24	R2 X MR	1.00	R2 X MR	0.95	R2 X MR	1.00
25	R3 X MR	-0.63	R3 X MR	-0.33	R3 X MR	0.54
26						
27	Low Correlation: 0.1-0.3		Med Correlation: 0.3-0.5		High Correlation: >0.5	
28	Same for Negative Correlations too...					

Can be done in R too…

Pearson product-moment correlation coefficient

From Wikipedia, the free encyclopedia

In statistics, the **Pearson product-moment correlation coefficient** (/ˈpɪərsɨn/) (sometimes referred to as the **PPMCC** or **PCC** or **Pearson's *r***) is a measure of the linear correlation (dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.[1][2][3]

Pearson product-moment correlation coefficient

Definition [edit]

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

For a population [edit]

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. The formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

where, cov is the covariance, σ_X is the standard deviation of X, μ_X is the mean of X, and E is the expectation.

For a sample [edit]

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter r and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for r by substituting estimates of the covariances and variances based on a sample into the formula above. That formula for r is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

An equivalent expression gives the correlation coefficient as the mean of the products of the standard scores. Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

where

$$\frac{X_i - \bar{X}}{s_X}, \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ and } s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

are the standard score, sample mean, and sample standard deviation, respectively.

Cohen's kappa

From Wikipedia, the free encyclopedia

Cohen's kappa coefficient is a statistical measure of inter-rater agreement or *inter-annotator agreement*[1] for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Some researchers[2][*citation needed*] have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can have the effect of underestimating agreement for a category that is also commonly used; for this reason, κ is considered an overly conservative measure of agreement.

Calculation [edit]

Cohen's kappa measures the agreement between two raters who each classify *N* items into *C* mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892),[4] see Smeeton (1985).[5]

The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where Pr(*a*) is the relative observed agreement among raters, and Pr(*e*) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(*e*)), κ = 0.

Agreement: Cohen's Kappa

Definition 1: If p_a = the proportion of observations in agreement and p_ε = the proportion in agreement due to chance, then **Cohen's kappa** is

$$\kappa = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon}$$

Alternatively

$$\kappa = \frac{n_a - n_\varepsilon}{n - n_\varepsilon}$$

where n = number of subjects, n_a = number of agreements and n_ε = number of agreements due to chance.

Agreement: Cohen's Kappa

PEARSON | =(F32-F33)/(F29-F33)

	A	B	C	D	E	F
1	Cohen's Kappa		<50 = 0	>50 = 1		
2						
3			Rater-1	Rater-2	Rater-3	
4						
5	headline1	Anger	0	0	0	
6	headline2	Anger	1	1	0	
7	headline3	Anger	0	0	0	
8						
9	headline1	Surprise	1	1	0	
10	headline2	Surprise	0	0	0	
11	headline3	Surprise	0	0	0	
12						
13	headline1	Disgust	0	0	1	
14	headline2	Disgust	0	0	1	
15	headline3	Disgust	1	1	1	
16		...				
17						
18						
19			Rater-1	Rater-2	Rater-3	
20	headline1		Surprise	Surprise	Disgust	
21	headline2		Anger	Anger	Disgust	
22	headline3		Disgust	Disgust	Disgust	
23						
24				Rater-2		
25			Anger	Surprise	Disgust	*Total*
26	Rater-1	Anger	1	0	0	1
27		Surprise	0	1	0	1
28		Disgust	0	0	1	1
29		*Total*	1	1	1	3
30						
31			Anger	Surprise	Disgust	*Total*
32		Agreements	1	1	1	3
33		Chance	0.33	0.33	0.33	1.00
34						
35		Kappa	/(F29-F33)			
36		Kappa	1			
37		Poor Agreeemnt = < 0.7		OK Agreement = > 0.7		
38						

Rater1 X Rater2

C35 | =(F32-F33)/(F29-F33)

	A	B	C	D	E	F
1	Cohen's Kappa					
2						
3			Rater-1	Rater-2	Rater-3	
4						
5	headline1	Anger	0	0	0	
6	headline2	Anger	1	1	0	
7	headline3	Anger	0	0	0	
8						
9	headline1	Surprise	1	1	0	
10	headline2	Surprise	0	0	0	
11	headline3	Surprise	0	0	0	
12						
13	headline1	Disgust	0	0	1	
14	headline2	Disgust	0	0	1	
15	headline3	Disgust	1	1	1	
16		...				
17						
18						
19			Rater-1	Rater-2	Rater-3	
20	headline1		Surprise	Surprise	Disgust	
21	headline2		Anger	Anger	Disgust	
22	headline3		Disgust	Disgust	Disgust	
23						
24				Rater-3		
25			Anger	Surprise	Disgust	*Total*
26	Rater-1	Anger	0	0	1	1
27		Surprise	0	0	1	1
28		Disgust	0	0	1	1
29		*Total*	0	0	3	3
30						
31			Anger	Surprise	Disgust	*Total*
32		Agreements	0	0	1	1
33		Chance	0.00	0.00	1.00	1
34						
35		Kappa	0.00000			
36		Kappa	0			
37						

Rater1 X Rater3

Agreement: Cohen's Kappa

- Note, there are also statistical methods for systematically detecting bias between groups of judges
- For example, that one judge has a tendency to prefer a certain category more than another (fitting is used)
- Some have used iterative bias detection

Expert Ratings Are Seldom Used…

- Are only really used for ground-truth
- Could not realistically rate all the items in a million item corpus or in a Tweet stream
- But, it has been used to identify docs/phrases/features for sentiment classifiers

Non-Expert Ratings

- Scaleability of expert ratings means they are seldom used, except for small nos of items

- But, ESPGame (vonAhn) and Amazon's Mechanical Turk (AMT) Human Intelligence Tasks (HITs) raised poss. of using non-experts

- Create HIT to do sentiment judgements, submit and collect answers…

amazon mechanical turk
beta
Artificial Artificial Intelligence

What is Mechanical Turk?

Sign In or Register | FAQ

Welcome | HITs | Qualifications

75,484 HITs available now

Introduction | Dashboard | Status | Your Account

Create Your Own HITs

Search for HITs containing that pay at least $ 0.00 for which you are qualified GO

To help search for Steve Fossett click here, then click "Accept HIT". New users may find the tutorial here useful. Learn about the results here.

Complete simple tasks that people do better than computers. And, get paid for it. Learn more.

Choose from thousands of tasks, control when you work, and decide how much you earn.

Do you want to quickly and easily create your own HITs on Amazon Mechanical Turk? Create HITs now.

If you are a software developer and would like to learn more about using Amazon Mechanical Turk APIs, click

STEP 1

Find
Find HITs to work on

What is a HIT?
HIT stands for Human Intelligence Task. These are tasks that people are willing to pay you to complete. For example a HIT might ask: "Is there a pizza parlour in this photograph?" Typically these tasks

STEP 2

Finish
Work & submit your HIT

How do I work on a HIT?
Once you have chosen a HIT to complete, click the "Accept HIT" button to have it assigned to you.

Follow the instructions on how to complete the HIT and when you are

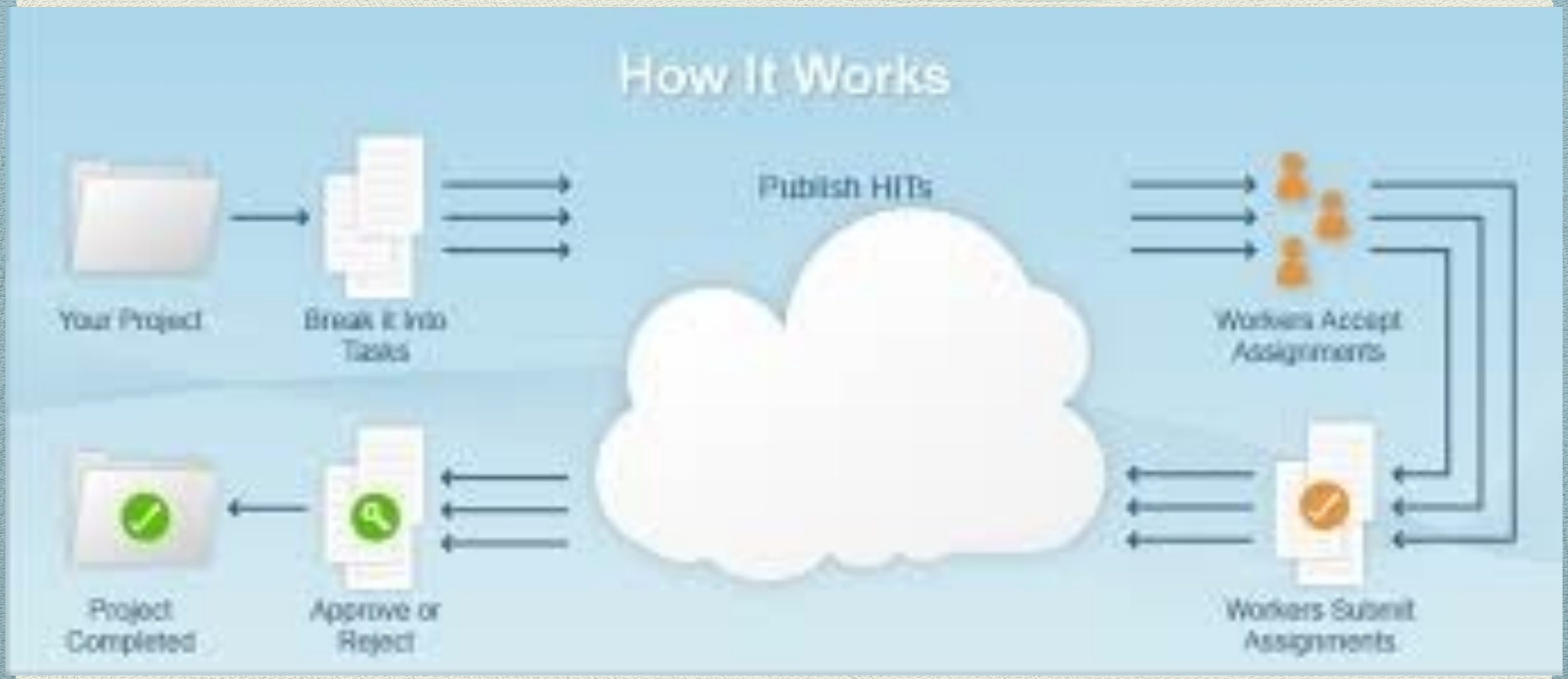
STEP 3

Earn
Get paid for your work

How do I get paid?
You are paid when your answer is approved by the person that listed the HIT.

The money you earn is deposited into your Amazon.com account, where

HITs in Mechanical Turk



HITs in Mechanical Turk

Add one complete sentence to the following story:

A lump rose in Andy's throat and he brushed away tears as he reread the chilling message taped to his dorm room door. This can't be happening he thought, How could someone do this to me? Andy carefully took the note off of the door, and entered his room.

- You must write a complete sentence that makes logical sense in the context of the story.
- Write the first sentence that comes to mind.
- Do not copy the sentence from another source.
- Begin with a asterisk character (*) if you would like to start a new paragraph.

Finished with this HIT? Let someone else do it?

Submit HIT Return HIT

Automatically accept the next HIT

Non-Experts & Experts

- Re-did S&M Affective Text Task; annotating headlines with emotions + valency
- Then looked at correlations to S&M's experts
- Found non-experts less good, but estimated no of non-experts to reach expert level

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast:but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254-263). ACL.

Snow et al (2008)…

Emotion	E vs. E	E vs. All	NE vs. E	NE vs. All
Anger	0.459	0.503	0.444	0.573
Disgust	0.583	0.594	0.537	0.647
Fear	0.711	0.683	0.418	0.498
Joy	0.596	0.585	0.340	0.421
Sadness	0.645	0.650	0.563	0.651
Surprise	0.464	0.463	0.201	0.225
Valence	0.759	0.767	0.530	0.554
Avg. Emo	0.576	0.603	0.417	0.503
Avg. All	0.580	0.607	0.433	0.510

Table 1: Average expert and non-expert ITA on test-set

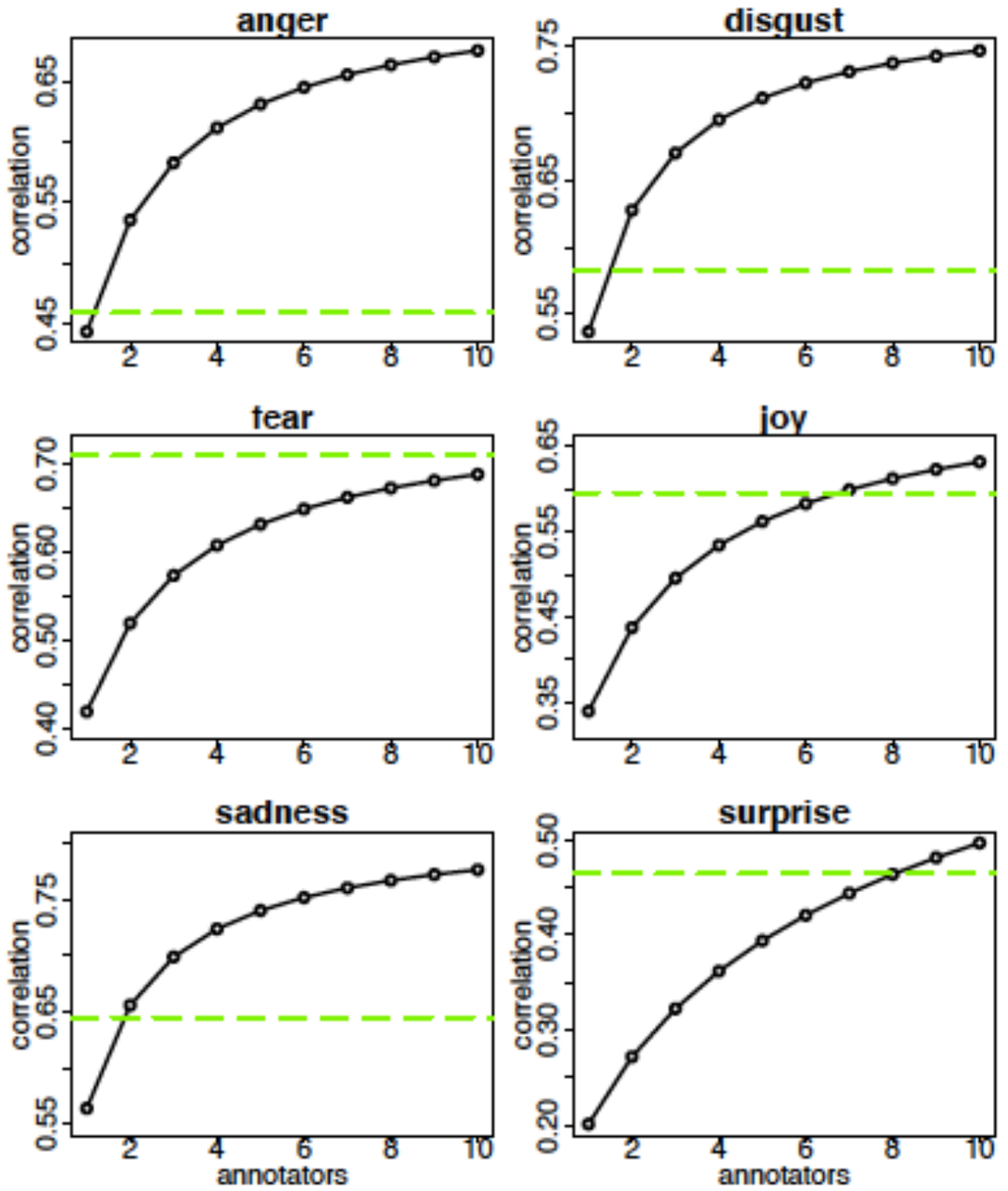


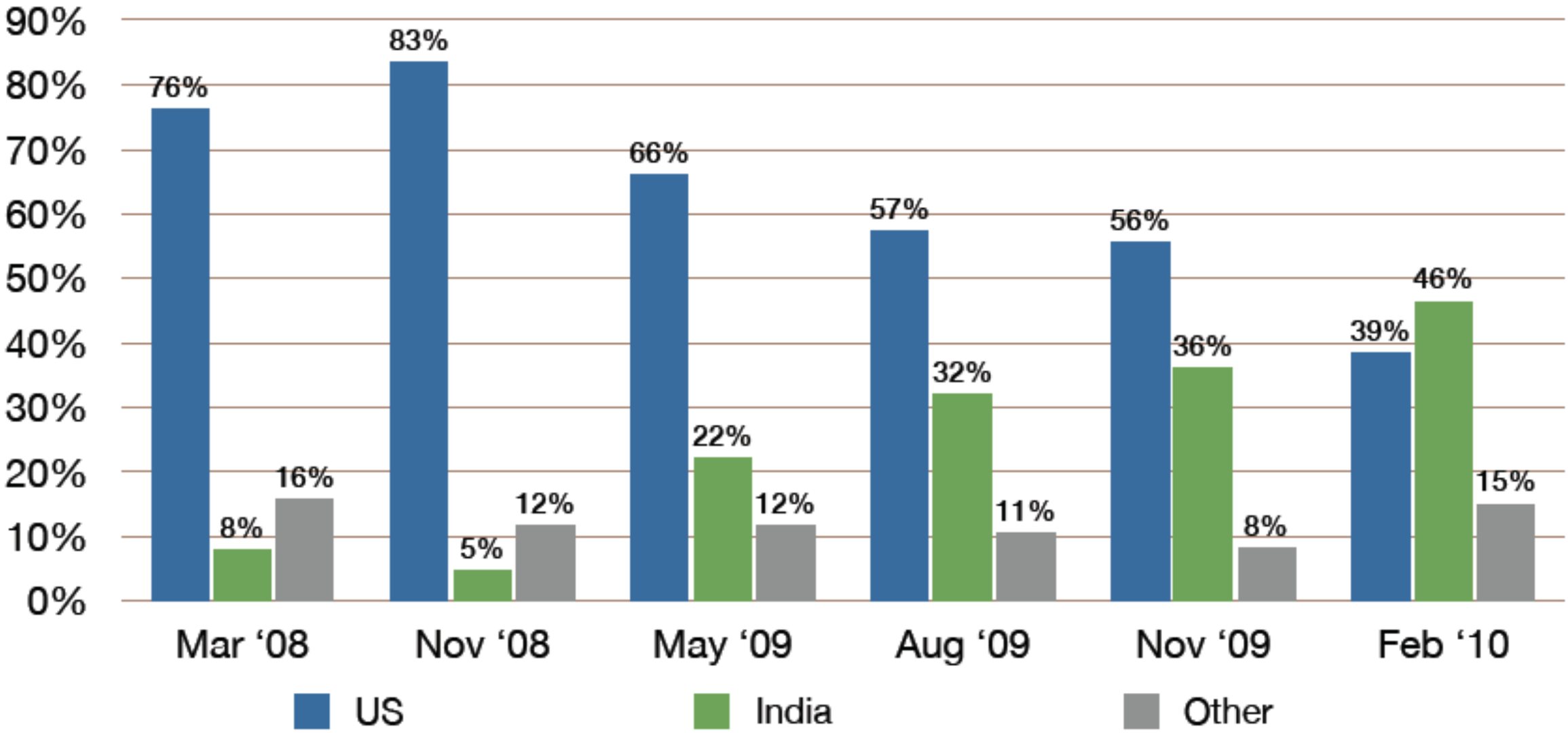
Figure 1: Non-expert correlation for affect recognition

Snow et al…

Emotion	1-Expert	10-NE	k	k-NE
Anger	0.459	0.675	2	0.536
Disgust	0.583	0.746	2	0.627
Fear	0.711	0.689	–	–
Joy	0.596	0.632	7	0.600
Sadness	0.645	0.776	2	0.656
Surprise	0.464	0.496	9	0.481
Valence	0.759	0.844	5	0.803
Avg. Emo.	0.576	0.669	4	0.589
Avg. All	0.603	0.694	4	0.613

Table 2: Average expert and averaged correlation over 10 non-experts on test-set. k is the minimum number of non-experts needed to beat an average expert.

But, see issues



Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010, April). Who are the crowdworkers? In CHI'10 (pp. 2863-2872). ACM.

Human Ratings: Conclusions

- Human ratings are not practically useful for large-scale, regular sentiment identification
- But, expert and non-expert ratings can form the basis for other systems; future of HITs unclear (moral)
- Indeed, they lie behind Sentiment Lexicons and Classifier solutions
- Methods for rater reliability are also generally important and useful

Ways to Identify Sentiments

- *Human Ratings*: of sentiment terms / phrases / documents

- ***Sentiment Lexicons:*** Matching text bits to sentiment word lists and lexicons (built from human judgements)

- ***Sentiment Classifiers:*** Classifying texts to find sentiment features (built from human ratings)

REM

Identifying Sentiment Terms:

Sentiment Lists & Lexicons

Sentiment Lexicons

- Another way to go is to develop lists of sentiment words, to build a lexicon
- Since, 1960s, social sciences have build such lexicons (nb. fall-stocks/fall-inflation issue)
- Examine three most commonly used ones

Identifying Using Lists

- Human (Basic) Emotions: anger, disgust, fear, happiness, sadness, and surprise

- Human Emotion Dimensions: Osgood, Suci & Tannenbaum (1965) semantic differentials

- Emotion word taxonomies: Johnson-Laird & Oatley (1989), Ortony, Clore & Foss (1987)

- Capture these emotions and their valency in lists

Egs of Lists

- Harvard General Inquirer List
- MPQA Corpus
- WordNet: Affective Wordnet, SentiWordNet

Harvard List: What it is?

- Lexicon attaching syntactic, semantic, and pragmatic information to part-of-speech tagged words

- Harvard IV-4 Dictionary of 13,000 words indicating syntax, positive, negative

- Very early market research, content analysis

Stone, P.J., Dunphry D.C., Smith, M.S., & Ogilvie, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Harvard List: What it is?

Table 3 A fragment of the Harvard General Inquirer spreadsheet file.

	Entry	Positiv	Negativ	Hostile	...184 classes ...	Othtags	Defined
1	A					DET ART	...
2	ABANDON		Negativ			SUPV	
3	ABANDONMENT		Negativ			Noun	
4	ABATE		Negativ			SUPV	
5	ABATEMENT					Noun	
...							
35	ABSENT#1		Negativ			Modif	
36	ABSENT#2					SUPV	
...							
11788	ZONE					Noun	

http://www.wjh.harvard.edu/~inquirer/

http://sentiment.christopherpotts.net/lexicons.html#opinionlexicon

Harvard List: Pro/Con

- Pro: fairly large lexicon, proved to be reasonable in identifying positive/negative texts (suicide notes)
- Con: still fairly minimal
- Con: superseded a bit, in recent years

MPQA Corpus

- MPQA Corpus: Multi-Perspective Question Answering, a Subjectivity Lexicon

- 15,991 subjective expressions (9k sentences from 425 docs) hand annotated using complex scheme; at word, phrase and sentence level

- Divided (words) into strong-subj and weak-subj and classified with polarity (partly use Harvard !)

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation, 39(2-3)*, 165-210.

Annotation Scheme

```
Direct subjective:
    Span: said
    Source: <writer,xirao-nima>
    Strength: high
    Expression strength: neutral
    Attitude type: negative
    Target: US
```

"The United States was slandering China again," said Xirao-Nima, a professor of Tibetan history at the Central University for Nationalities." [Beijing China Daily. Mar 2002]

```
Direct subjective:
    Span:  slandering
    Source: <writer, xirao-nima,US>
    Target: China
    Strength: high
    Expression strength: high
```

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation, 39(2-3)*, 165-210.

Agreement…

To measure the reliability of the polarity annotation scheme, we conducted an agreement study with two annotators, using 10 documents from the MPQA Corpus. The 10 documents contain 447 subjective expressions. Table 1 shows the contingency table for the two annotators' judgments. Overall agreement is 82%, with a Kappa (κ) value of 0.72.

	Neutral	Positive	Negative	Both	Total
Neutral	123	14	24	0	161
Positive	16	73	5	2	96
Negative	14	2	167	1	184
Both	0	3	0	3	6
Total	153	92	196	6	447

Table 1: Agreement for Subjective Expressions (Agreement: 82%, κ: 0.72)

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). ACL.

MPQA: What it is?

A fragment of the MPQA subjectivity lexicon.

Strength	Length	Word	Part-of-speech	Stemmed	Polarity
type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
type=strongsubj	len=1	word1=zest	pos1=noun	stemmed1=n	priorpolarity=positive

http://mpqa.cs.pitt.edu/

http://sentiment.christopherpotts.net/lexicons.html#opinionlexicon

Parse Rules + Classify

Word Features	Sentence Features	Structure Features
word token	strongsubj clues in current sentence: count	in subject: binary
word part-of-speech	strongsubj clues in previous sentence: count	in copular: binary
word context	strongsubj clues in next sentence: count	in passive: binary
prior polarity: positive, negative, both, neutral	weaksubj clues in current sentence: count	
reliability class: strongsubj or weaksubj	weaksubj clues in previous sentence: count	
Modification Features	weaksubj clues in next sentence: count	**Document Feature**
preceeded by adjective: binary	adjectives in sentence: count	document topic
preceeded by adverb (other than not): binary	adverbs in sentence (other than not): count	
preceeded by intensifier: binary	cardinal number in sentence: binary	
is intensifier: binary	pronoun in sentence: binary	
modifies strongsubj: binary	modal in sentence (other than will): binary	
modifies weaksubj: binary		
modified by strongsubj: binary		
modified by weaksubj: binary		

Table 3: Features for neutral-polar classification

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354). ACL.

MPQA

- Pro: Scaleable and commonly used
- Pro: Works at word, phrase, sentence levels
- Pro: Combines syntax and lexicon
- Con: Validation looks a bit tricky
- Con: Statistical basis for combining ratings

What is Wordnet?

- G.A. Miller (1995) founded WordNet as lexical resource (big names work on it)
- High quality lexical resource, for nouns, verbs, adjs; does not define meaning of words but POS and synsets
- 117659 synsets, 115k words, 200k word-sense pairs organised by conceptual relations between synsets (super-subordinate, partonymy)

http://wordnet.princeton.edu/

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*: 39-41.

Wordnet Synset

- Unordered synonym sets (car / automobile;close / shut)
- Gloss of meaning (not definition)
- Example of use (as in OED)
- Set of relations between these synsets ; car ISA Vehicle, vehicle ISA artefact, artefact ISA object

http://wordnet.princeton.edu/

Miller, G.A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41. or Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Synset Eg

gloss

usage

ISA

ISA

linked data reference

```
<Synset id="eng-30-06645039-n" baseConcept="1">
    <Definition gloss="mark of a foot or shoe on a surface">
      <Statement example="the police made casts of the footprints
                          in the soft earth outside the window" />
    </Definition>
    <SynsetRelations>
<!-- {mark, print}: -->
     <SynsetRelation target="eng-30-06798750-n" relType="has_hyperonym" >
      <Meta author="AH" date="2008-07-01" source="Wordnet3.0"
            status="yes" confidenceScore="1.0" />
     </SynsetRelation>
<!-- {footprint_evidence}: -->
     <SynsetRelation target="eng-30-06645266-n" relType="has_hyponym" >
      <Meta author="AH" date="2008-07-01" source="eng-Wordnet3.0"
            status="yes" confidenceScore="1.0" />
    </SynsetRelation>
    </SynsetRelations>
    <MonolingualExternalRefs>
      <MonolingualExternalRef
          externalSystem="SUMO"
          externalReference="superficialPart" relType="at"/>
    </MonolingualExternalRefs>
</Synset>
```

Synset Hierarchy / Ontology?

```
dog, domestic dog, Canis familiaris
    => canine, canid
       => carnivore
          => placental, placental mammal, eutherian, eutherian mammal
             => mammal
                => vertebrate, craniate
                   => chordate
                      => animal, animate being, beast, brute, creature, fauna
                         => ...
```

http://en.wikipedia.org/wiki/WordNet

Synset Relations

Most synonym sets are connected to other synsets via a number of semantic relations. These relations vary based on the type of word, and include:

- Nouns
  - *hypernyms*: *Y* is a hypernym of *X* if every *X* is a (kind of) *Y* (*canine* is a hypernym of *dog*)
  - *hyponyms*: *Y* is a hyponym of *X* if every *Y* is a (kind of) *X* (*dog* is a hyponym of *canine*)
  - *coordinate terms*: *Y* is a coordinate term of *X* if *X* and *Y* share a hypernym (*wolf* is a coordinate term of *dog*, and *dog* is a coordinate term of *wolf*)
  - *meronym*: *Y* is a meronym of *X* if *Y* is a part of *X* (*window* is a meronym of *building*)
  - *holonym*: *Y* is a holonym of *X* if *X* is a part of *Y* (*building* is a holonym of *window*)
- Verbs
  - *hypernym*: the verb *Y* is a hypernym of the verb *X* if the activity *X* is a (kind of) *Y* (*to perceive* is an hypernym of *to listen*)
  - *troponym*: the verb *Y* is a troponym of the verb *X* if the activity *Y* is doing *X* in some manner (*to lisp* is a troponym of *to talk*)
  - *entailment*: the verb *Y* is entailed by *X* if by doing *X* you must be doing *Y* (*to sleep* is entailed by *to snore*)
  - *coordinate terms*: those verbs sharing a common hypernym (*to lisp* and *to yell*)
- Adjectives
  - *related nouns*
  - *similar to*
  - *participle of verb*
- Adverbs
  - *root adjectives*

http://en.wikipedia.org/wiki/WordNet

WordNet Use…

- Can be used in lots of different ways; traverse hierarchy to determine similarity?
- Expand queries with synset items, can improve retrieval or conceptual processing
- Glosses/egs can be further analysed
- WordNet conferences/open source community/ API in nltk

What is WordNet Affect

- Selected and tagged sentiment words from WordNet with: 4.8k words, 2.8k synsets
- Added an Affect Slot, to identify core emotion category —>
- Added A-label to synsets
- nb. not polarity

Category	Example Term
emotion	anger
cognitive state	doubt
trait	competitive
behaviour	cry
attitude	skepticism
feeling	pleasure

Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In LREC (Vol. 4, pp. 1083-1086).

WordNet Affect Egs

A-Labels	Examples
EMOTION	noun anger#1, verb fear#1
MOOD	noun animosisy#1, adjective amiable#1
TRAIT	noun aggressiveness#1, adjective competitive#1
COGNITIVE STATE	noun confusion#2, adjective dazed#2
PHYSICAL STATE	noun illness#1, adjective all_in#1
EDONIC SIGNAL	noun hurt#3, noun suffering#4
EMOTION-ELICITING SITUATION	noun awkwardness#3, adjective out_of_danger#1
EMOTIONAL RESPONSE	noun cold_sweat#1, verb tremble#2
BEHAVIOUR	noun offense#1, adjective inhibited#1
ATTITUDE	noun intolerance#1, noun defensive#1
SENSATION	noun coldness#1, verb feel#3

Table 4: A-Labels and corresponding example synsets

Strapparava, C., & Valitutti, A. (2004, May). WordNet Affect: an Affective Extension of WordNet. In LREC (Vol. 4, pp. 1083-1086).

What is SensiWordNet

- Add Polarity to WordNet terms: *Obj(s), Pos(s), Neg(s);* assign scores for each, add to 1
- **Estimable**(1): *may be estimated;* Obj(1), Pos(0), Neg(0)
- **Estimable**(2): *deserving of respect;* Obj(.25), Pos(.75), Neg(0)
- Scores are derived from processing glosses (pre-processed and tf-idf-ed) through an ensemble of different classifiers as pos/neg/obj

Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, 6, 417-422.

SensiWordNet

- Find largish no of words of sentiment
- More in adjectives
- Evaluation not complete

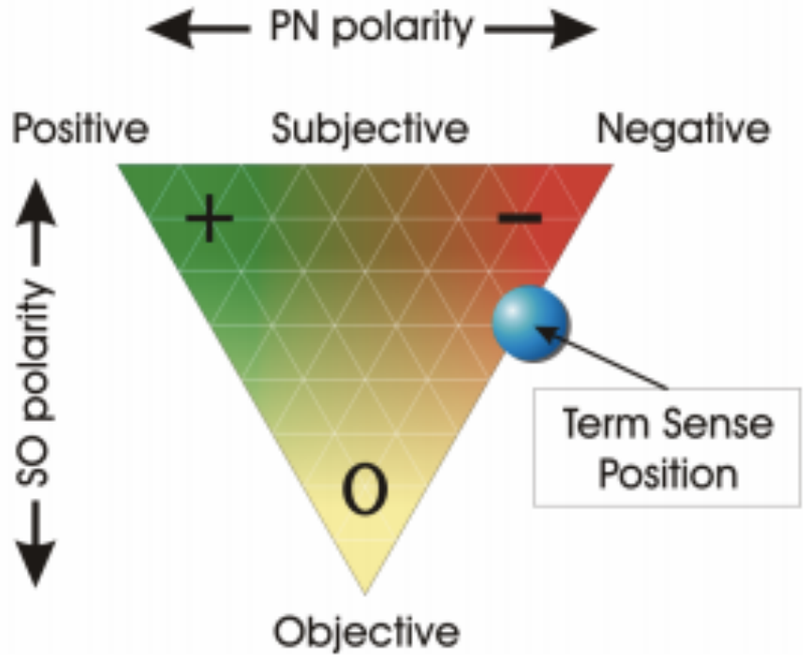
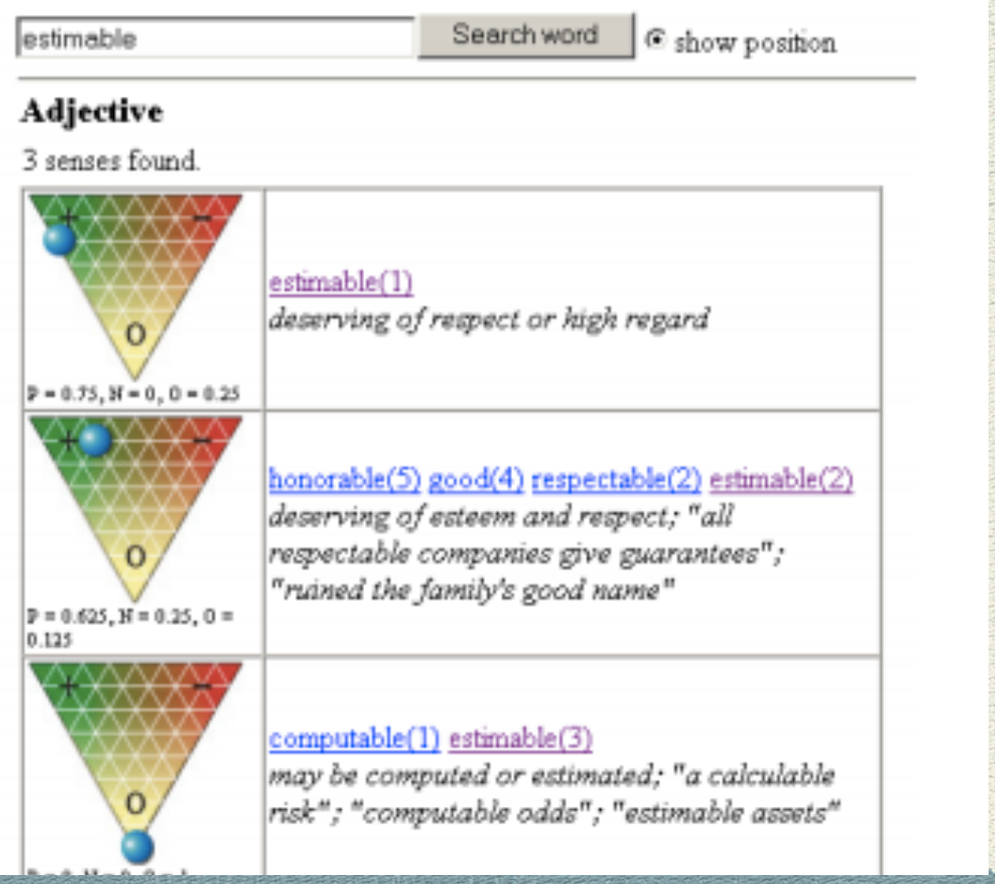


Figure 1: The graphical representation adopted by SENTIWORDNET for representing the opinion-related properties of a term sense.



Lexicons & Lists: Pros/Cons

- Pro: Plugs affective analysis into very large lexicon, used by many systems
- Pro: Seems to solve a lot of the problems
- Con: Unreliability of human ratings and/or automated extensions of lists
- Con: Inherent +/- of terms, modified in context ; “its not a bad film”

Ways to Identify Sentiments

- *Human Ratings*: of sentiment terms / phrases / documents

- ***Sentiment Lexicons:*** Matching text bits to sentiment word lists and lexicons (built from human judgements)

- ***Sentiment Classifiers:*** Classifying texts to find sentiment features (built from human ratings)

REM

Identifying Sentiment Terms:

Classifying & Extracting Features

Two Ways to Use Classifiers

- ***Unsupervised***: use probabilistic word-similarity to known positive/negative words (eg "excellent", "poor") to classify a doc as positive or negative

- ***Supervised***: train a classifier on known positive or negative docs and extract pos/neg features from it (using n-Bayes, SVM)

Unsupervised Classification

- Turney (2002) takes doc as input and outputs that doc recommends / no-recommends item

- Extract phrases, determine their similarity to pos / negative words using PMI (SO)

- Average the SO score for all phrases in doc

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of 40th ACL (pp. 417-424).

Unsupervised: PMI

Pointwise mutual information

The PMI of a pair of outcomes *x* and *y* belonging to discrete random variables *X* and *Y* quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. Mathematically:

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

The mutual information (MI) of the random variables *X* and *Y* is the expected value of the PMI over all possible outcomes (with respect to the joint distribution $p(x,y)$).

The measure is symmetric ($\operatorname{pmi}(x;y) = \operatorname{pmi}(y;x)$). It can take positive or negative values, but is zero if *X* and *Y* are independent. PMI maximizes when *X* and *Y* are perfectly associated, yielding the following bounds:

$$-\infty \leq \operatorname{pmi}(x;y) \leq \min\left[-\log p(x), -\log p(y)\right].$$

Finally, $\operatorname{pmi}(x;y)$ will increase if $p(x|y)$ is fixed but $p(x)$decreases.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist., 16*, 22–29.

PMI

$$PMI(word_1, word_2) = \log_2 \left[\frac{p(word_1 \ \& \ word_2)}{p(word_1)\, p(word_2)} \right] \quad (1)$$

- $p(word_1 \ \& \ word_2)$ is probability of joint occurrence of words in a phrase
- $p(word_1)p(word_2)$ if statistically independent, probability of co-occurrence is the product
- ratio is a measure of the statistical independence of the two words
- $\log_2$ of ratio = amount of information that we get about the presence of one word when we observe the other

Turney, P. D. (2002). Thumbs up or thumbs down?. *ACL*, 417-424.

SO

$$SO(phrase) = PMI(phrase, \text{“excellent”}) - PMI(phrase, \text{“poor”}) \quad (2)$$

- Compute the Semantic Orientation (SO) by seeing if PMI with positive word is greater than PMI with negative word

- For probability of “excellent” / “poor” use Altavista Hits (i.e., hits for “Rocky-II”, hits for “Rocky-II excellent”, hits for “Rocky-II poor”)

- Average the SO scores over the document, if + then its a recommend, if - a no-recommend

Turney, P. D. (2002). Thumbs up or thumbs down?. *ACL*, 417-424.

Outputs: PMI-IR

Table 2. An example of the processing of a review that the author has classified as *recommended*.[6]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		0.322

Table 3. An example of the processing of a review that the author has classified as *not recommended*.

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

$$\text{SO}(\textit{phrase}) = \log_2 \left(\frac{\text{hits}(\text{phrase NEAR “excellent”}) \, \text{hits}(\text{“poor”})}{\text{hits}(\text{phrase NEAR “poor”}) \, \text{hits}(\text{“excellent”})} \right) \quad (1)$$

Turney, P. D. (2002). Thumbs up or thumbs down? *ACL*, 417-424.

Unsupervised; Results

Table 6. The confusion matrix for movie classifications.

	Author's Classification		
Average Semantic Orientation	Thumbs Up	Thumbs Down	Sum
Positive	28.33 %	12.50 %	40.83 %
Negative	21.67 %	37.50 %	59.17 %
Sum	50.00 %	50.00 %	100.00 %

Turney, P. D. (2002). Thumbs up or thumbs down? *ACL*, 417-424.

Supervised Classifiers

- Pang et al (2002) train a classifier on known positive or negative docs and extract pos/neg features from it, using NaiveBayes and SVM
- IMDb movie database of reviews; converting star ratings into pos/neg/neutral
- Limited 20 reviews per category per author; 144 reviews; 752 negative, 1301 positive

Pang et al (2002)

- Pit human-extracted words (69% over 50% chance) against several classifiers
- Naive Bayes, Maximum Entropy, SVM (L7)
- Uses bag-of-words, each doc is a vector of features based on frequency of occurrence

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *ACL-02 Empirical Methods in Natural Language Processing,10* (pp. 79-86). ACL

Pang et al (2002): Setup

- Randomly selected 700 pos and 700 neg
- Created 3 equal-sized folds from these (for training and testing)
- Pre-processing: removed punctuation, but not stemming or stop-word removal
- Explored unigrams, bigrams, POS, using frequency or presence (1/0)

Pang et al (2002): Results

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

Figure 3: Average three-fold cross-validation accuracies, in percent. Boldface: best performance for a given setting (row). Recall that our baseline results ranged from 50% to 69%.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *ACL-02 Empirical Methods in Natural Language Processing,10* (pp. 79-86). ACL

Pang et al (2002): Conclusions

- Classification did better than humans
- Presence better than frequencies (often noted)
- SVM seemed like best method, but not by a lot
- Sentiment classification harder than topic clssf.
- POS tagging does not generate big gains

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *ACL-02 Empirical Methods in Natural Language Processing,10* (pp. 79-86). ACL

Supervised Classifiers

- Pang et al (2002) trained classifier on docs; can also train on sentences

- Kim & Hovy (2004) trained Naive Bayes on sentences with identified polarities

- Trying to get at subtleties at sentence level, though short of full NLP this is hard

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. In *20th Int. Conf. on Comp. Ling.* (p. 1367). ACL

Supervised V Unsupervised

- Pit Turney's PMI against Pang et al Classifiers
- Find that overall classifiers do quite well but that for real-time systems unsupervised might be better; training takes long time

Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. HICSS'05, pp. 112c. IEEE.

Supervised Classifiers: Pros/Cons

- Are often the preferred solution to problems
- But, there are issues:
  - Learning takes a long time
  - What features have been learned ?
  - Generalisation to other areas, often poor; drop to chance outside movie area

Supervised Classifiers: Learn?

- Sometimes the features learned seem weird:
  - in movie set, neg reviews mention director, plot, writer script; pos mention ending, parts
  - as, *yet*, with, both occur more in positive
  - since, *have*, those, though occur more in negative
  - Some words are negative only in the movie context; video, tv, series

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, 267-307.

Mixed Methods

- Also note, you will find many cases that mix these various techniques up and down; a lexicon with a classifier, PMI with a dictionary…this always happens

Ways to Identify Sentiments

- *Human Ratings*: of sentiment terms / phrases / documents

- ***Sentiment Lexicons:*** Matching text bits to sentiment word lists and lexicons (built from human judgements)

- ***Sentiment Classifiers:*** Classifying texts to find sentiment features (built from human ratings)

REM

Liu's Quintuple Stands

An *opinion* is a quintuple

$(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$,

where

- e_j is a target entity.
- a_{jk} is an aspect/feature of the entity e_j.
- so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l. so_{ijkl} is +ve, -ve, or neu, or more granular ratings.
- h_i is an opinion holder.
- t_l is the time when the opinion is expressed.

Solution not complete

Factual Descriptions can seem negative

Sarcasm is a problem

How short of full NLP can you go?

Table 8: Examples of problems causing misclassification

First Example	*Embedded factual information*
Sample Phrase:	Horrific sequence [adjective noun]
SO of Sample Phrase:	-1.2547 (negative)
Context of Sample Phrase:	In my opinion, in what is probably the films most **horrific sequence**, Francis travels to a nearby insane asylum to check if anyone by the name of Caligari has been there or escaped.
Author's rating:	A+ (positive)
Movie:	Cabinet of Dr. Caligari
Second Example	*Sarcastic review writing*
Sample Phrase:	Terrifically written [adverb adjective]
SO of Sample Phrase:	0.0455 (positive)
Context of Sample Phrase:	As with most highly depressing movies, the story in About Schmidt dragged very slowly for the first 100% of it (the last 0% was actually very well paced and **terrifically written**).
Author's rating:	D- (negative)
Movie:	About Schmidt

Conclusions I

- Different options for identifying sentiment
  - Human ratings, Lists and Lexicons
  - Unsupervised classification
  - Supervised classification
- Leaves us with the issue (largely) of how you use methods to achieve various tasks (e.g. prediction)

Conclusions II

- Note what Liu says about difficulties;
- Issue is really how much NLP is required to achieve results

Overview

- Identifying Sentiment Terms (Lect-9)
- Using Sentiments to Do Things (Lect-10)
- Key Examples (both)
- Some Implementations (both)