



Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai


Presenter: Liangchen Fan
Hauptseminar: Mining Volatile Data
Dr. Eirini Ntoutsi
Wintersemester 2012/13

[http://www.dbs.ifi.lmu.de/cms/Hauptseminar_\"Mining_Volatile_Data\"_WS1213](http://www.dbs.ifi.lmu.de/cms/Hauptseminar_\)

- Outline

- I Introduction
- II Definitions
- III Topic-Sentiment Mixture Model
- IV Experiments and Results
- V Conclusion
- VI Discussion

- What is „sentiment“ and „sentiment analysis“?

sen·ti·ment /'sentəmənt/ 

Noun: 1. A view of or attitude toward a situation or event; an opinion.
2. General feeling or opinion: "racist sentiment".

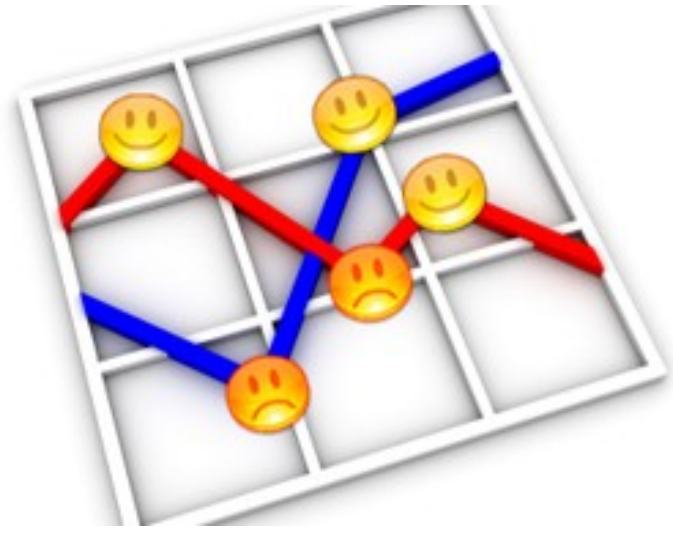
Synonyms: feeling - sense - opinion - emotion - sensation

 or  ?



A legend for sentiment analysis with five categories, each with a corresponding emoji icon:

-  Positive
-  Somewhat Positive
-  Neutral
-  Somewhat Negative
-  Negative



- Related Work

- Sentiment Classification
- Topic Modelling

- This Work

- Mixture of Topic and Sentiment Research
- Subtopics
 - a set of documents: $C = \{d_1, d_2, \dots, d_m\}$
 - k major topics(subtopics) with topic model $\theta: \{\theta_1, \theta_2, \dots, \theta_k\}$
- Sentiment Polarities: positive, negative
- Problem: Topic-Sentiment Analysis(TSA)
- Model: Topic-Sentiment Mixture Model(TSM)

- What is „topic-sentiment analysis“?

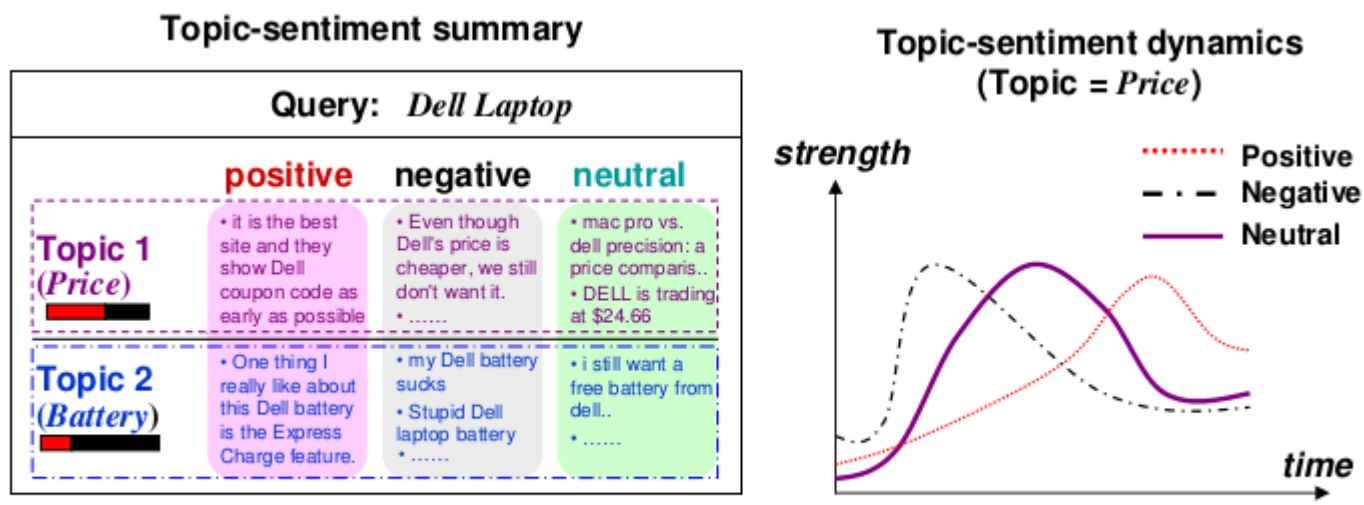


Figure 1: A possible application of topic-sentiment analysis

• Definitions

- - Topic Model:

- probabilistic distribution of word w in a set of words V : $\{p(w|\theta)\}_{w \in V}$

- words coherence: $\sum_{w \in V} p(w|\theta) = 1$

- - Sentiment Model

- positive words: $(\{p(w|\theta_p)\}_{w \in V}) \quad \sum_{w \in V} p(w|\theta_p) = 1$

- negative words: $(\{p(w|\theta_N)\}_{w \in V}) \quad \sum_{w \in V} p(w|\theta_N) = 1$

- - Sentiment Coverage of topic θ_i in document d :

- coverage of positive opinons: $\delta_{i,d,P}$

- coverage of neutral opinons: $\delta_{i,d,F}$

- coverage of negative opinions: $\delta_{i,d,N}$

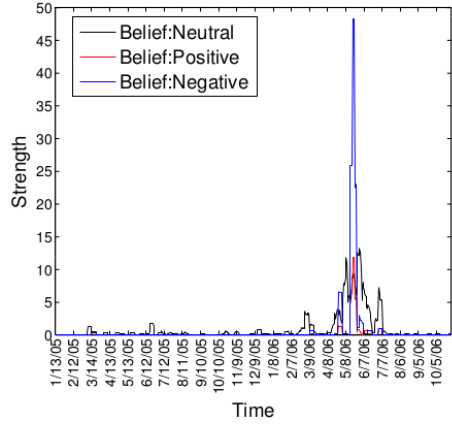
$$C_{i,d} = \{\delta_{i,d,F}, \delta_{i,d,P}, \delta_{i,d,N}\}$$

$$\delta_{i,d,F} + \delta_{i,d,P} + \delta_{i,d,N} = 1$$

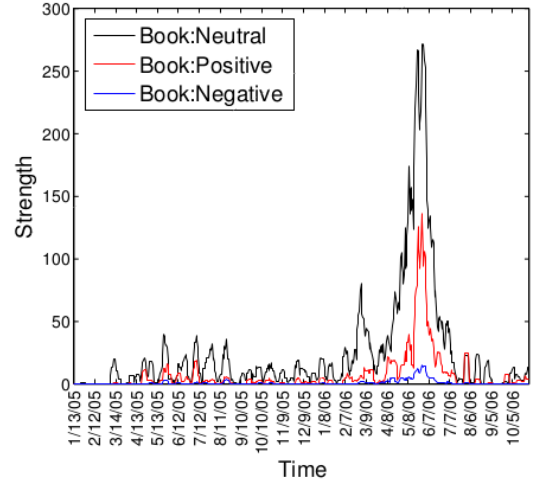
- Definitions

- Topic Life Cycle:
 - the amount of document content about a topic over time

- Sentiment Dynamics:
 - the distribution of the strength(the amount) of a sentiment(positive, negative, neutral) about a topic over time



(b) Da Vinci Code: Religion



(a) Da Vinci Code: Book

- Topic-Sentiment Analysis(TSA)
 - Learning General Sentiment Models
 - Extracting Topic Models and Sentiment Coverages
 - Modeling Topic Life Cycle and Sentiment Dynamics

- The Generation Process

- Categories of Words

- common english words: „the“, „a“, „of“
 - words related to a topical theme: „iPad“, „price“, „screen“
 - the two categories above are captured with a background component model
 - for words related to a topic: three subcategories
 - neutral opinions: „price“, „screen“
 - positive opinions: „awesome“, „love“
 - negative opinions: „bad“, „hate“

- The Generation Process
 - Multinomial Models
 - Background Model θ_B
 - a set of topic models: $\{\theta_1, \theta_2, \dots, \theta_k\}$
 - Positive sentiment model: θ_P
 - Negative sentiment model: θ_N
 - How does the author choose a word while expressing an opinion?
(Example: Opinions about iPadmini)

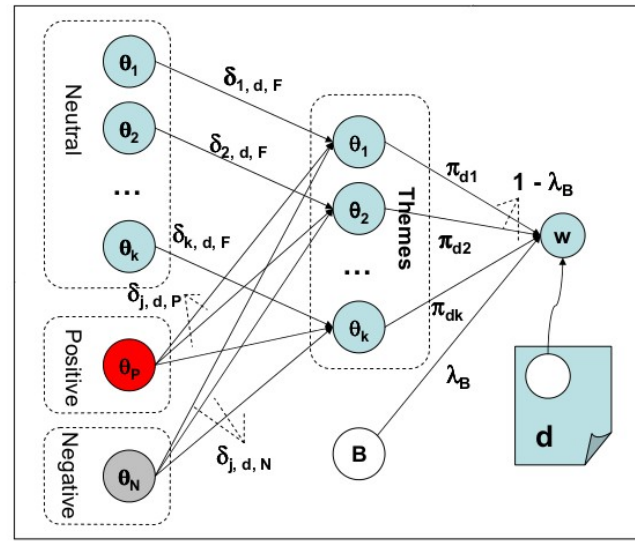


Figure 2: The generation process of the topic-sentiment mixture model

Topic Sentiment Mixture

> III Topic-Sentiment Mixture Model

- Topic-Sentiment Mixture Model

- Formular
- C: a collection of weblog articles

$$\log(\mathcal{C}) = \sum_{d \in \mathcal{C}} \sum_{w \in V} c(w : d) \log \left[\lambda_B p(w|B) + (1 - \lambda_B) \sum_{j=1}^k \pi_{dj} \times \left(\delta_{j,d,F} p(w|\theta_j) + \delta_{j,d,P} p(w|\theta_P) + \delta_{j,d,N} p(w|\theta_N) \right) \right]$$

Background Model
neutral positive negative

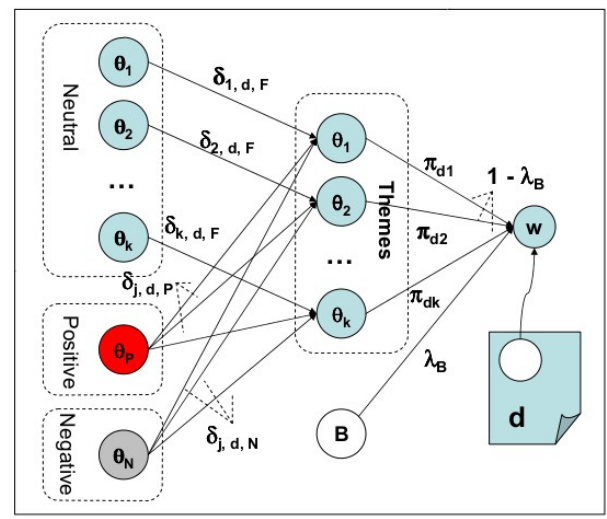


Figure 2: The generation process of the topic-sentiment mixture model

- Topic-Sentiment Mixture Model
 - Two-Step Estimation Framework
 - The first step: defining model priors:
 - to tell the TSM what the sentiment models should look like
 - Opinmind: retrieves online sentiments
 - When given a query(topic), it can retrieve positive and negative sentences(with sentiment labels)
 - mixture of results retrieved with various queries

- Topic-Sentiment Mixture Model
 - Two-Step Estimation Framework
 - The second step: maximizing a posterior estimation:
 - Expectation-Maximization Algorithm(EM)
 - based on the prior $p(\Lambda)$
 - M-Step Updates with the MAP estimator: $\hat{\Lambda} = \arg \max_{\Lambda} p(C|\Lambda)p(\Lambda)$
 - in positive, negative and neural sentiment models

- Further Tasks

- based on the EM Algorithm and the M-step Updates
- Rank Sentences for Topics

$$Score_j(s) = -D(\theta_j || \theta_s) = - \sum_{w \in V} p(w|\theta_j) \log \frac{p(w|\theta_j)}{p(w|\theta_s)}$$

- Categorize Sentences by Sentiments

$$\arg \max_x -D(\theta_s || \theta_x) = \arg \max_x - \sum_{w \in V} p(w|\theta_s) \log \frac{p(w|\theta_s)}{p(w|\theta_x)}$$

- Reveal the Overall Opinions for documents/Topics

$$S(j, P) = \frac{\sum_{d \in \mathcal{C}} \pi_{dj} \delta_{j,d,P}}{\sum_{d \in \mathcal{C}} \pi_{dj}}$$

- Sentiment Dynamics Analysis
 - the change of the positive and negative opinions about a topic over time
 - understand the public opinions better
 - predict user behavior more accurate
 - Hidden Markov Model(HMM)
 - Tag every word in the collection with a topic and sentiment polarity, but which sentiment word is about which topic can't be tagged
 - an alternative HMM:
 - Count the Words with Corresponding Labels Over Time (based the first presentation)

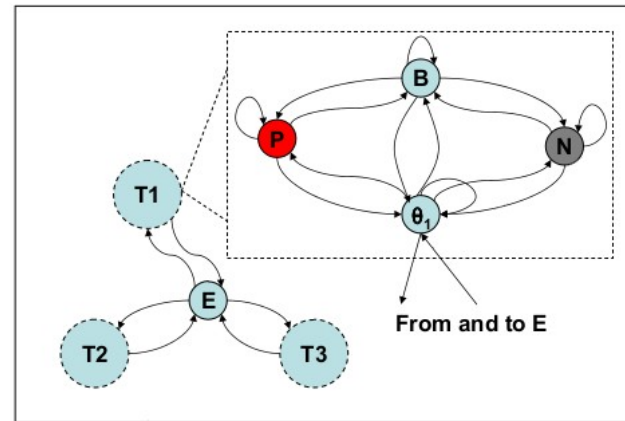


Figure 4: The Hidden Markov Model to extract topic life cycles and sentiment dynamics

- Data Sets

- One Set for Learning: a dataset retrieved with Opinmind

Topic	# Pos.	# Neg.	Topic	# Pos.	# Neg.
laptops	346	142	people	441	475
movies	396	398	banks	292	229
universities	464	414	insurances	354	297
airlines	283	400	nba teams	262	191
cities	500	500	cars	399	334

Table 1: Basic statistics of the OPIN data sets

- The Other Set for Testing: a „raw“ dataset with timestamp about topics „iPod“ and „Da Vinci Code“

Data Set	# doc.	Time Period	Query Term
iPod	2988	1/11/05~11/01/06	ipod
Da Vinci Code	1000	1/26/05~10/31/06	da+vinci+code

Table 2: Basic statistics of the TEST data sets

- Sentiment Model Extraction
 - Differences bt. General and Biased Sentiment Models
 - Unbiased sentiment models contain neutral contents.

P-mix	N-mix	P-movies	N-movies	P-cities	N-cities
love	suck	love	hate	beautiful	hate
awesome	hate	harry	harry	love	suck
good	stupid	pot	pot	awesome	people
miss	ass	brokeback	mountain	amaze	traffic
amaze	fuck	mountain	brokeback	live	drive
pretty	horrible	awesome	suck	good	fuck
job	shitty	book	evil	night	stink
god	crappy	beautiful	movie	nice	move
yeah	terrible	good	gay	time	weather
bless	people	watch	bore	air	city
excellent	evil	series	fear	greatest	transport

Table 3: Sentiment models learnt from a mixture of topics are more general

- Topic Model Extraction
 - Without Prior: more confused
 - With Prior: more informative and coherent

NO-Prior			With-Prior	
batt., nano	marketing	ads, spam	Nano	Battery
battery	apple	free	nano	battery
shuffle	microsoft	sign	color	shuffle
charge	market	offer	thin	charge
nano	zune	freepay	hold	usb
dock	device	complete	model	hour
itunes	company	virus	4gb	mini
usb	consumer	freeipod	dock	life
hour	sale	trial	inch	rechargeable

Table 4: Example topic models with TSM: iPod

- Topic Model Extraction
 - Without Prior: more confused
 - With Prior: more informative and coherent

NO-Prior			With-Prior	
content	book	background	movie	religion
langdon	author	jesus	movie	religion
secret	idea	mary	hank	belief
murder	holy	gospel	tom	cardinal
louvre	court	magdalene	film	fashion
thrill	brown	testament	watch	conflict
clue	blood	gnostic	howard	metaphor
neveu	copyright	constantine	ron	complaint
curator	publish	bible	actor	communism

Table 5: Example topic models: Da Vinci Code

- Topic Model Extraction
 - Result of query „da vinci code“ by Opinmind

	Neutral	Thumbs Up	Thumbs Down
Topic1 (Movie)	... Ron Howards selection of Tom Hanks to play Robert Langdon.	Tom Hanks stars in the movie, who can be mad at that?	But the movie might get delayed and even killed off if he loses.
	Directed by: Ron Howard Writing credits: Akiva Goldsman ...	Tom Hanks, who is my favorite movie star act the leading role.	protesting ... will lose your faith by ... watching the movie
	After watching the movie I went online and some research on ...	Anybody is interested in it?	... so sick of people making such a big deal about a FICTION book and movie.
Topic2 (Book)	I knew this because I was once a follower of feminism.	And I'm hoping for a good book too.	... so sick of people making such a big deal about a FICTION book and movie.
	I remembered when i first read the book, I finished the book in two days.	Awesome book.	This controversy book cause lots conflict in west society.
	I'm reading "Da Vinci Code" now.	So still a good book to past time.	in the feeling of deeply anxious and fear, to ... read books calmly was quite difficult.

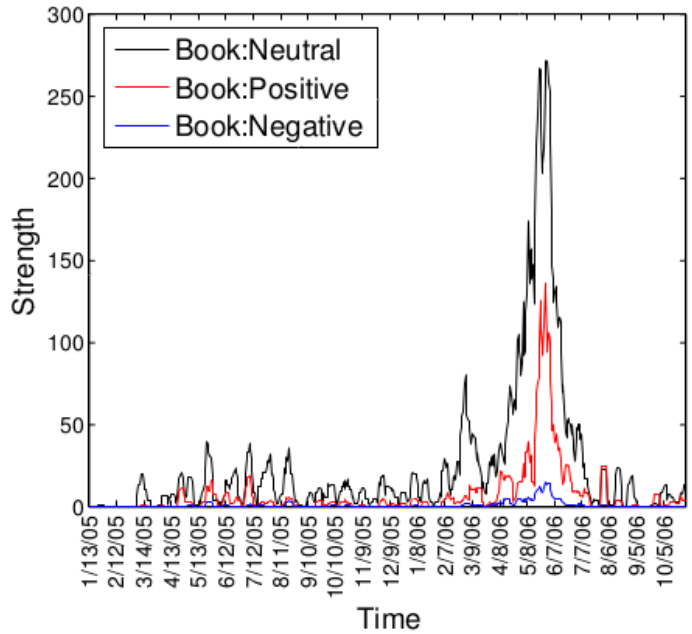
Table 6: Topic-sentiment summarization: Da Vinci Code

- Topic Model Extraction
 - iPod

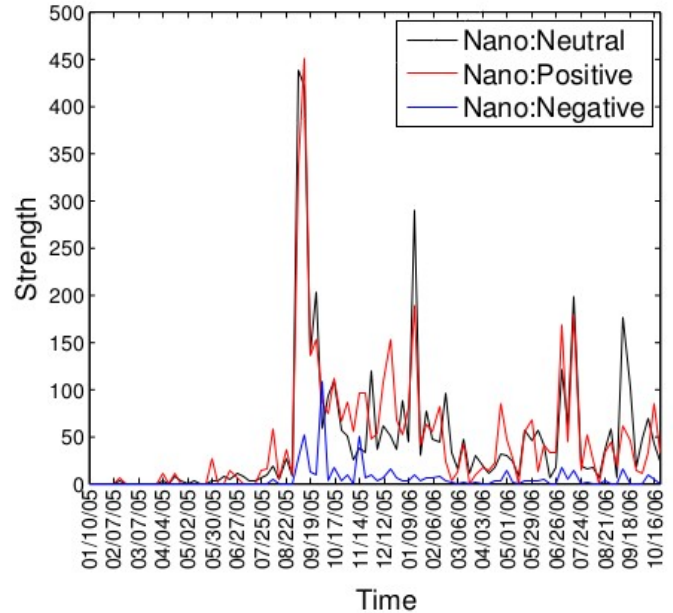
TSM		
	Thumbs Up	Thumbs Down
1	(sweat) iPod Nano ok so ... Ipod Nano is a cool design, ...	WAT IS THIS SHIT??!! ipod nanos are TOO small!!!!
2	the battery is one serious example of excellent relibability	Poor battery lifeiPod's battery completely died
3	My new VIDEO ipod arrived!!! Oh yeah! New iPod video	fake video ipod Watch video podcasts ...

Opinmind	
Thumbs Up	Thumbs Down
I love my iPod, I love my G5...	I hate ipod.
I love my little black 60GB iPod	Stupid ipod out of batteries...
I LOVE MY IPOD	" hate ipod " = 489..
I love my iPod.	my iPod looked uglier...surface...
- I love my iPod.	i hate my ipod.
... iPod video looks SO awesome	... microsoft ... the iPod sucks

- Topic Life Cycle and Sentiment Dynamics
 - Time Axis: 2005-01-13 to 2006-10-05, interval 30 days
 - Single Burst vs Continuous Burst



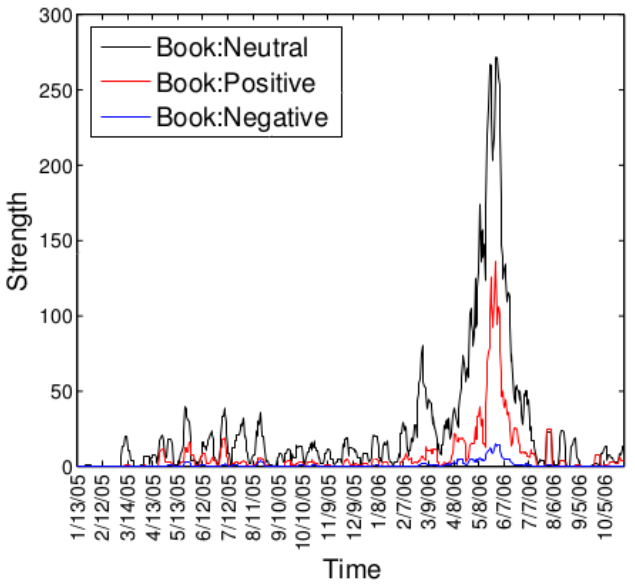
(a) Da Vinci Code: Book



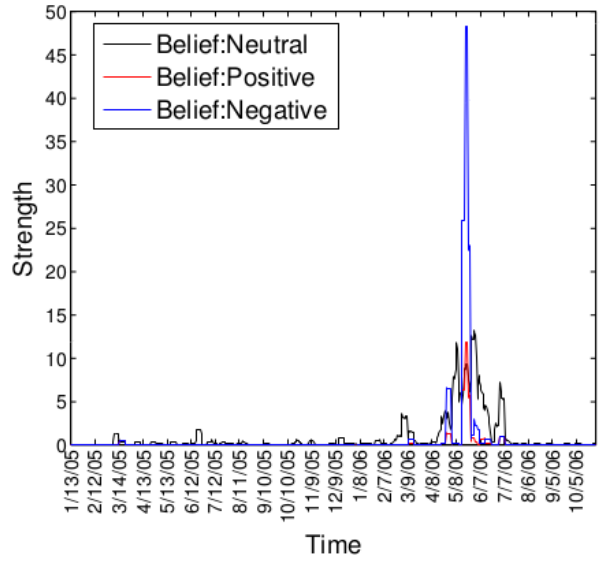
(c) iPod: Nano

• Topic Life Cycle and Sentiment Dynamics

- Burst in 2006-05 in „Book“ and „Religion“ because of the movie, in „Book “ from 2006-04
- In „Book“: more positive opinions
- In „Religion“: more negative opinions

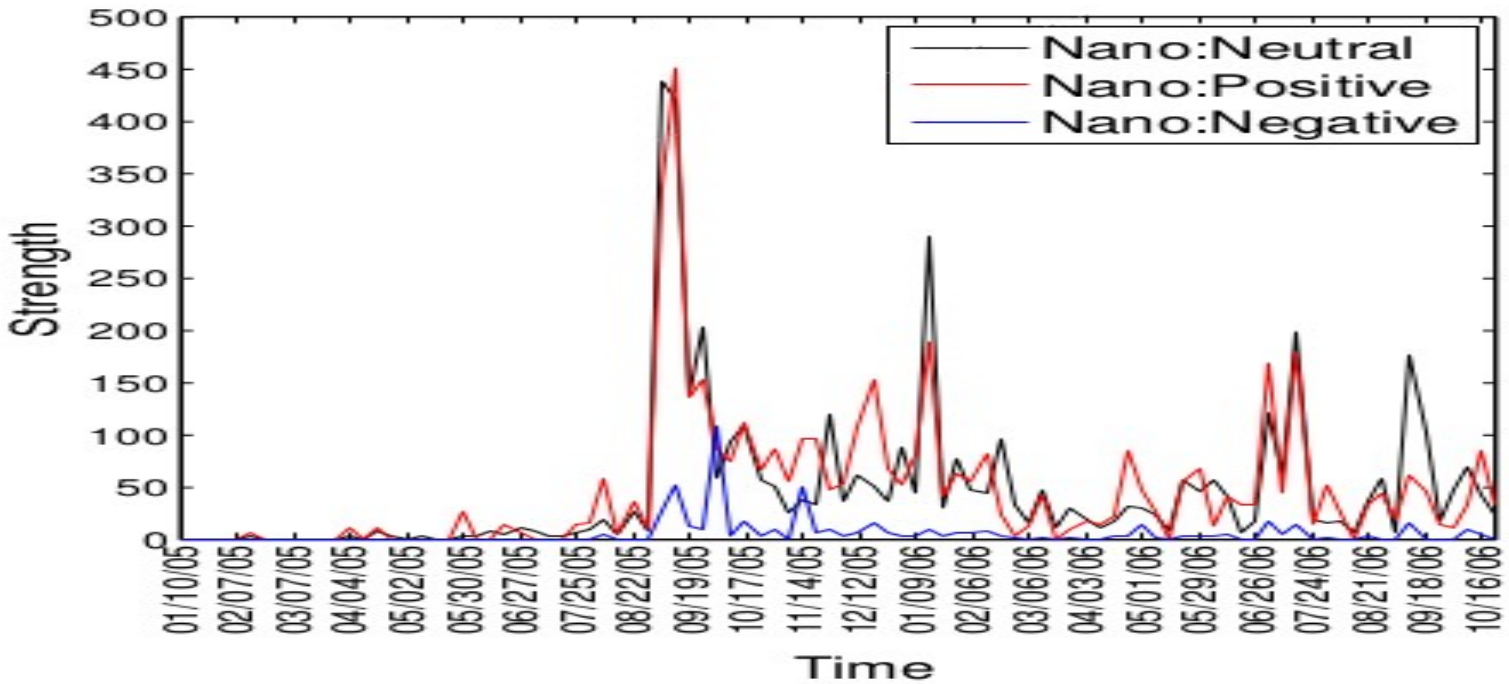


(a) Da Vinci Code: Book



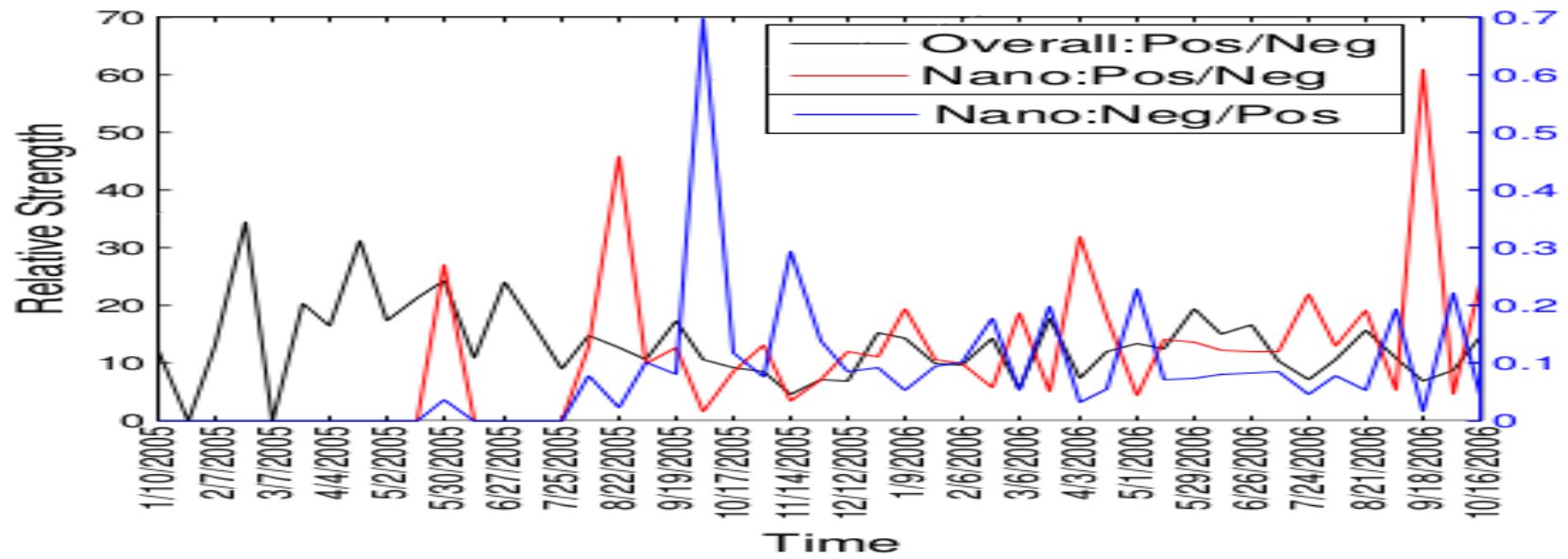
(b) Da Vinci Code: Religion

- Topic Life Cycle and Sentiment Dynamics
 - In 2005-09 the burst of positive and neutral, later the burst of negative



(c) iPod: Nano

- Topic Life Cycle and Sentiment Dynamics
 - Relative: Pos/Neg and Neg/Pos with different scale
 - the sentiment domination and its trends
 - Blackline(Pos/Neg of „iPod“) is getting weaker



(d) iPod: Relative

- Topic-Sentiment Mixture Model(TSM)
- Further Development
 - Customize the sentiment models according to each topic and obtain different contextual views
 - user behavior prediction
- My Comments
 - The mixture model release more accurate sentiment mining results associated to many different topics in a topic-mixed documents. It is applicable to mega media datas, a very strong tool for „content analysis“.
 - It can be more accuracy on „subtopics“ of „subtopics“.

- Discussion



- Thank you!

*Thank
you*