

"Topic sentiment mixture: modeling facets and opinions in weblogs."

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

Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

Roadmap

- Introduction
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- Conclusion

Roadmap

- **Introduction**
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- Conclusion

Introduction

- Weblogs i.e., blogs
 - Wide coverage of topics
 - Dynamics of discussion
 - Abundance of Opinions
- Makes them extremely valuable for mining user opinions
- Analysis of blogs would enable applications like
 - Opinion search for ordinary users
 - Opinion tracking for business intelligence
 - User behavior prediction for targeted advertising

Introduction

Mining user opinions from Weblogs



Sentiment Analysis of Blog Data

Introduction

- Existing work on blog data
 - Extracting and analyzing topical content
 - Without any sentiment analysis
- Why is this a limitation?

Roadmap

- Introduction
- **Motivation**
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- Conclusion

Motivation

Let us consider an example of Amazon.com

Sales of a product \propto No. of blog posts

However,

~~*Sales of book \propto Burst of criticism*~~

Similarly, decrease of blog mentions about a product might be caused by a decrease in complaints about it

Motivation



- Understanding pos/neg about topic/subtopic important
- Some existing works do take sentiment into consideration in Weblogs but,
 - Only overall sentiment of a blog
 - What about sentiment regarding subtopic?
- Blogs often cover range of subtopics and may have different opinions about each.

Motivation

- For example,
 - User may like *fuel* and *efficiency* of a car but dislike its *power* and *safety* aspects
- People always have different opinions about subtopics
- General opinion about a topic is not that informative

Motivation

Topic-sentiment summary

Query: <i>Dell Laptop</i>			
	positive	negative	neutral
Topic 1 <i>(Price)</i> 	<ul style="list-style-type: none">• it is the best site and they show Dell coupon code as early as possible	<ul style="list-style-type: none">• Even though Dell's price is cheaper, we still don't want it.•	<ul style="list-style-type: none">• mac pro vs. dell precision: a price comparis..• DELL is trading at \$24.66
Topic 2 <i>(Battery)</i> 	<ul style="list-style-type: none">• One thing I really like about this Dell battery is the Express Charge feature.	<ul style="list-style-type: none">• my Dell battery sucks• Stupid Dell laptop battery•	<ul style="list-style-type: none">• i still want a free battery from dell..•

Roadmap

- Introduction
- Motivating Examples
- **Problem Formulation**
 - **Topic Sentiment Analysis problem**
 - **Problem Definition**
 - **Challenges**
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- Conclusion

Topic Sentiment Analysis Problem

Let, $C = \{d_1, d_2, \dots, d_m\}$ be set of documents

We assume that it covers number of topics, subtopics and related sentiments

Let us assume that there are k major topics in the documents each characterized by a unigram language model

$\theta_1, \theta_2, \dots, \theta_k$ be the k topic models

Topic Sentiment Analysis Problem

- Topic Model
 - $\{p(\omega|\theta)\}_{(\omega \in V)}$
 - $\sum_{\omega \in V} \{p(\omega|\theta)\}_{(\omega \in V)} = 1$
 - k such topic models
- Sentiment Models
 - $\{p(\omega|\theta_P)\}_{(\omega \in V)}, \{p(\omega|\theta_N)\}_{(\omega \in V)}$
 - $\sum_{\omega \in V} \{p(\omega|\theta_P)\}_{(\omega \in V)}, \sum_{\omega \in V} \{p(\omega|\theta_N)\}_{(\omega \in V)}$
 - Orthogonal to topic models
- Sentiment Coverage of a Topic
 - (Relative coverage of neutral, pos and neg opinions) / doc
 - $c_i, d = \{\delta_{(i,d,F)}, \delta_{(i,d,P)}, \delta_{(i,d,N)}\}, \delta_{(i,d,F)} + \delta_{(i,d,P)} + \delta_{(i,d,N)} = 1$

Topic Sentiment Analysis Problem

- Topic Life Cycle
 - Time series representing strength distribution of neutral contents of a topic over time
- Sentiment Dynamics
 - Time series representing strength distribution of pos/neg opinions about a topic over time

Problem Definition

- Learning general sentiment models
- Extracting topic models and sentiment coverages
 - Customize the general sentiment models learned previously for a given collection and extract topic models and sentiment coverage for each topic
- Model topic lifecycle and sentiment dynamics

Challenges

- Mixture modelling of topic and sentiment unclear
 - Existing topic extraction frameworks not able to extract sentiment models from text
 - Sentiment classification algorithm can't model mixture of topics simultaneously
- Sentiment classification algorithms tend to overfit training data

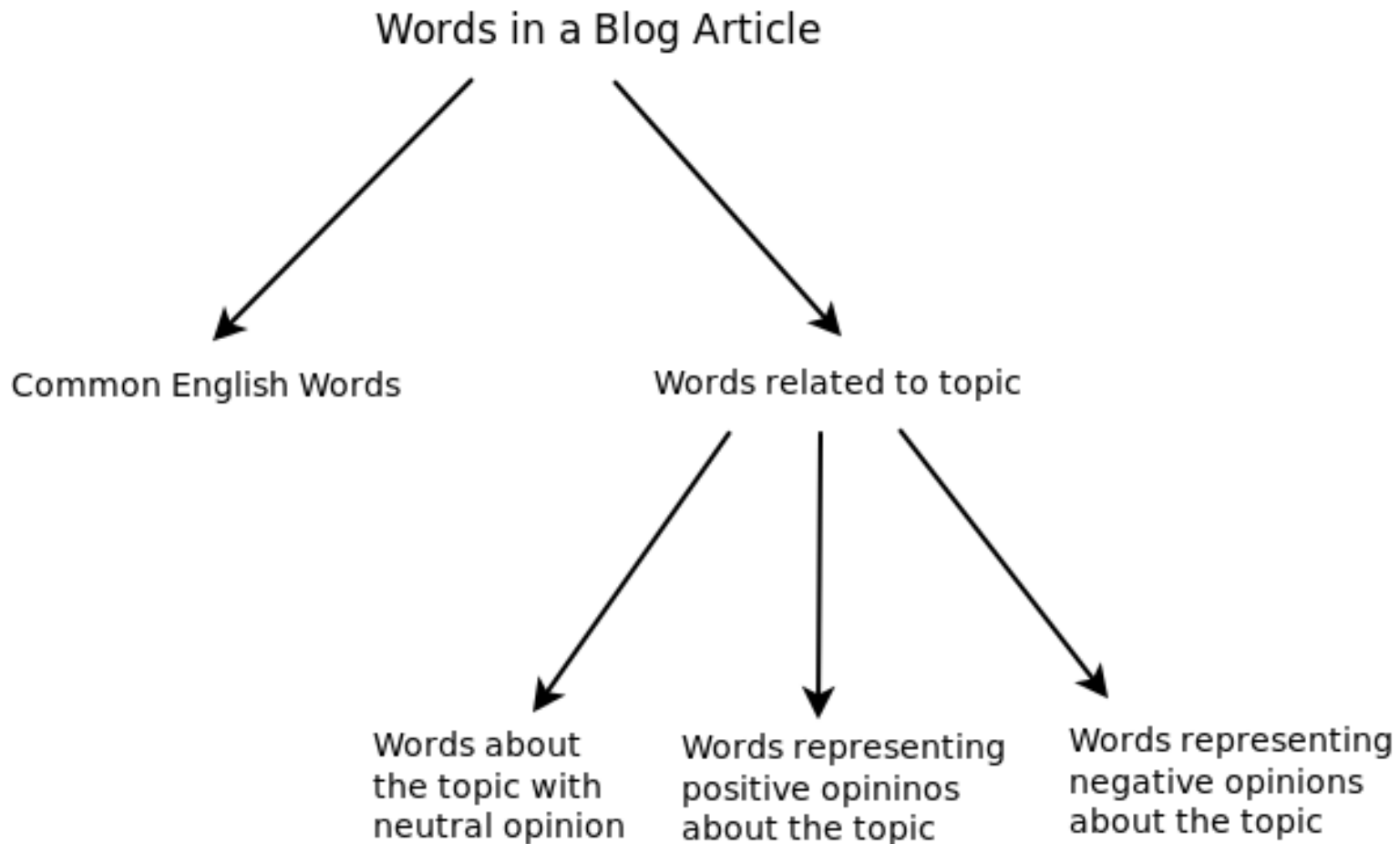
Roadmap

- Introduction
- Motivation
- Problem Formulation
- **Topic Sentiment Mixture**
 - Mixture of Language Models
 - Generation Process
 - TSM Model
 - Sentiment Model Priors
 - MAP Estimation
 - Utilizing the Model
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- Conclusion

Mixture of language models

- Mixture of multinomial distributions
 - Used for extraction of topics/subtopics
- None of the works models topics and sentiments simultaneously
- In this work, mixture of multinomials includes two sentiment models

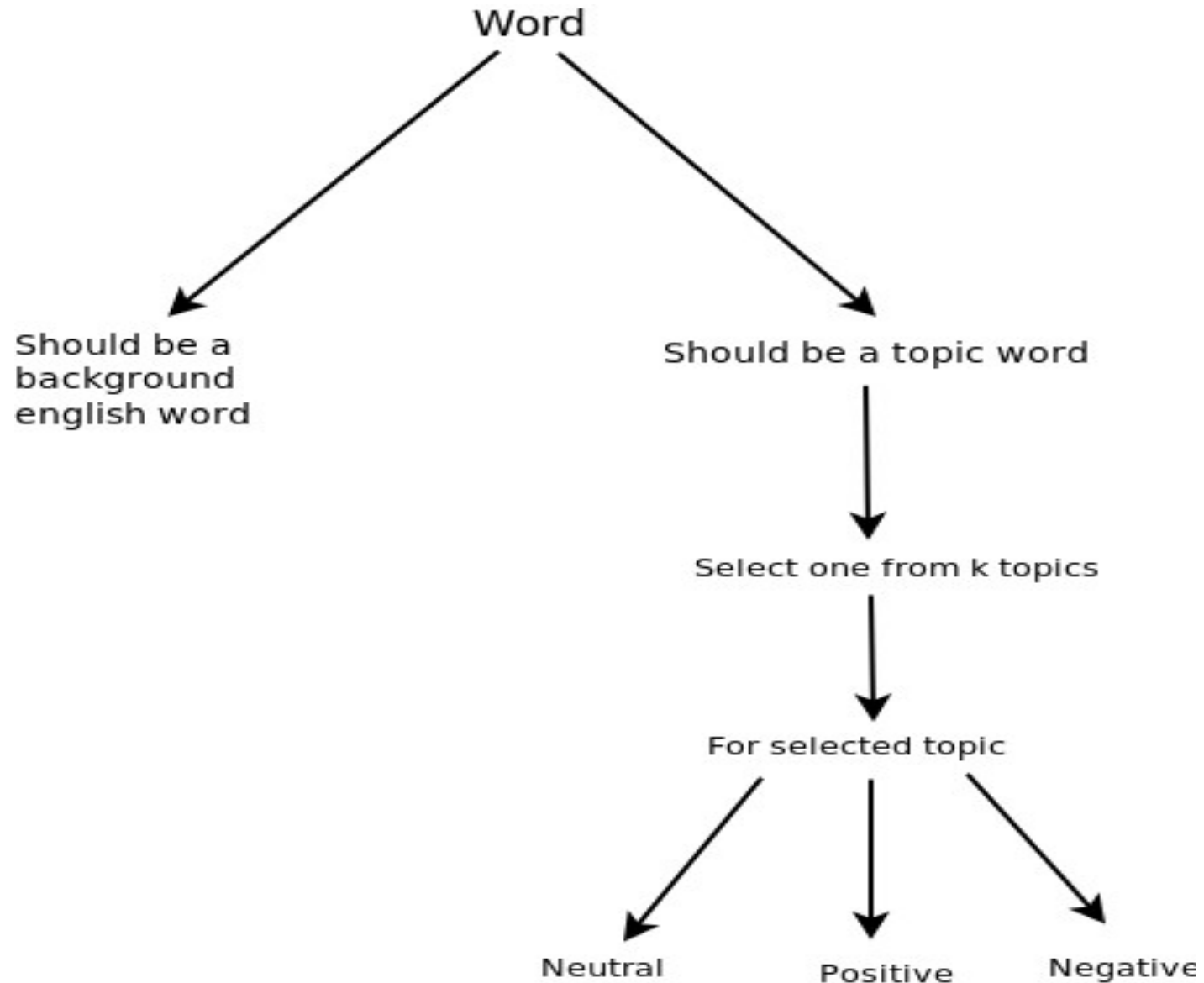
Mixture of language models



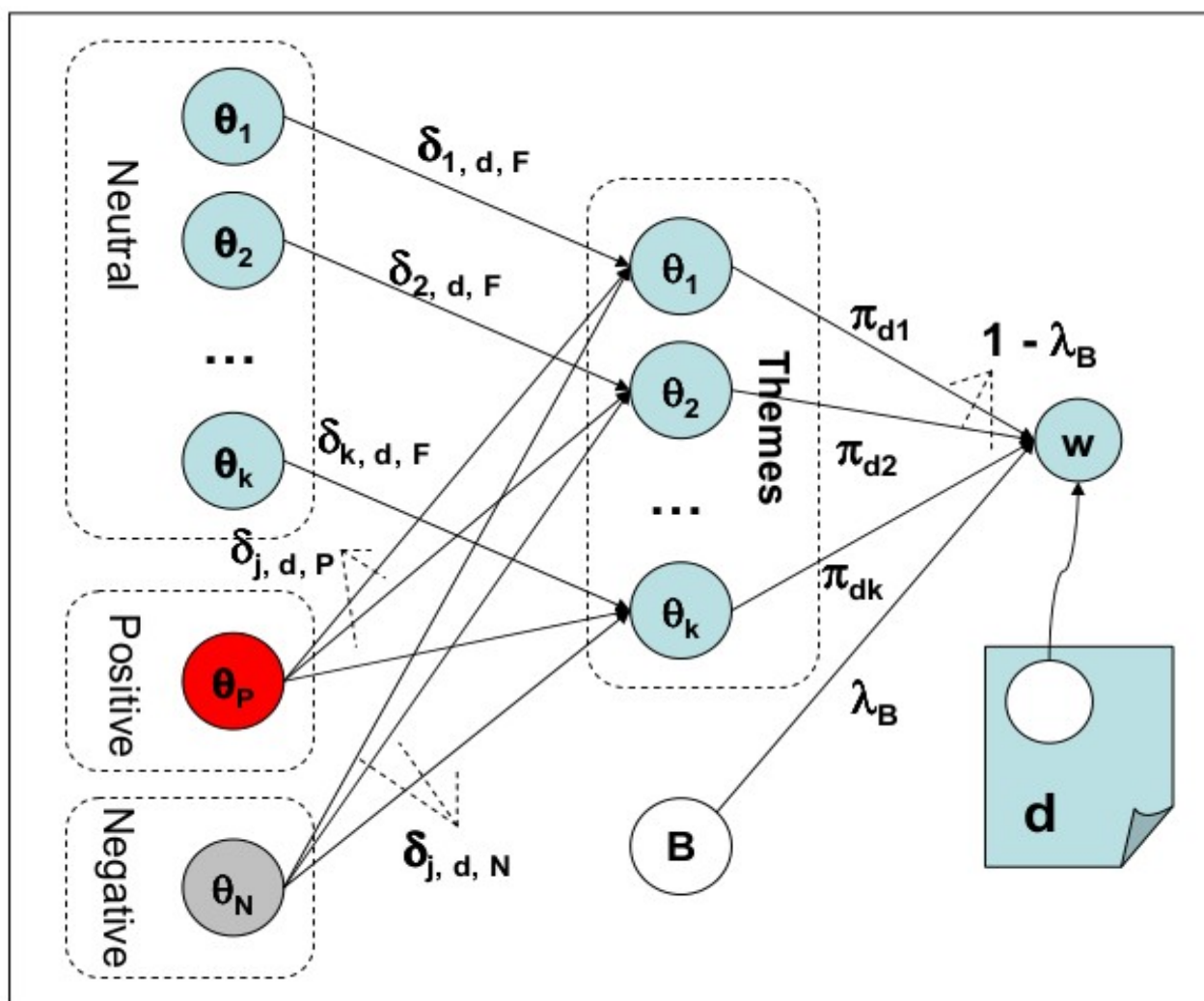
Mixture of language models

- Background model to capture common english models, θ_B
- k topic models, $\theta_1, \theta_2, \dots, \theta_k$
- Positive sentiment model, θ_P
- Negative sentiment model, θ_N

Generation Process



Generation Process



TSM Model

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

$c(w:d)$ = count of word w in document d

λ_B is the probability of choosing θ_B

π_{d_j} is the probability of choosing the j^{th} topic in document d

$\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$ is the sentiment coverage of topic j in document d

λ_B is set to an empirical constant between 0 and 1

TSM Model

- For background model,

$$p(\omega|\theta_B) = \frac{\sum_{d \in C} c(\omega:d)}{\sum_{\omega \in V} \sum_{d \in C} c(\omega \in d)}$$

of times word appears
In the collection



Word count of whole
collection

TSM Model

- Parameters remaining to be estimated
 - Topic models - $\theta_1, \theta_2, \dots, \theta_k$
 - Sentiment models - θ_P and θ_N
 - Document topic probabilities - π_{d_j}
 - Sentiment coverage for each document - $\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$
- Let us denote these free parameters by Λ

TSM Model

- ML estimation using EM algorithm
 - Sentiment models biased towards specific contents in the collection
 - Topic models will also give high probability to sentiment bearing words due to co-occurrence
- MAP estimation is used in this case
 - Learn Priors for sentiment models
 - Combine prior with the data likelihood to estimate parameters using MAP estimation

Sentiment Model Priors

- Model Prior
 - Sentiment model priors should tell TSM how the actual sentiment models look like in the working collection
 - Domain specific lexicons?
 - Not possible for every topic
 - Solution – general sentiment models
- Opinmind
 - Retrieves pos/neg sentences about a topic
 - Queries in various topics to ensure diversity
 - $C = \{(d, t_d, s_d)\}$, t_d indicates topic, s_d indicates sentiment
- TSM model is used to fit the training data and estimate sentiment models
 - $\pi_{d_j} = 1$ if $t_d = j$ and $\pi_{d_j} = 0$ otherwise
 - $\delta_{j,d,P} = 0$ if s_d is negative and $\delta_{j,d,N} = 0$ if s_d is positive

Sentiment Model Priors

- Let $\overline{\theta}_P$ and $\overline{\theta}_N$ be the learnt sentiment models
- $Dir(\{1 + \mu_P p(\omega | \overline{\theta}_P)\}_{\omega \in V})$ is the conjugate dirichlet prior for θ_P
- $Dir(\{1 + \mu_N p(\omega | \overline{\theta}_N)\}_{\omega \in V})$ is the conjugate dirichlet prior for θ_N
- Prior knowledge about the topic is used to define a topic prior

$$p(\Lambda) \propto p(\theta_P) * p(\theta_N) * \prod_{j=1}^k p(\theta_j)$$

$$p(\Lambda) = \prod_{\omega \in V} p(\omega | \theta_P)^{\mu_P p(\omega | \overline{\theta}_P)} \prod_{\omega \in V} p(\omega | \theta_N)^{\mu_N p(\omega | \overline{\theta}_N)} \prod_{j=1}^k \prod_{\omega \in V} p(\omega | \theta_j)^{\mu_j p(\omega | \overline{\theta}_j)}$$

MAP Estimation

$$\bar{\Lambda} = \underset{\Lambda}{\operatorname{argmax}} p(C|\Lambda) p(\Lambda)$$

Utilizing the model

- Rank sentences for topics

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

θ_s is the smoothed language model for sentence s

- Categorize sentences by sentiments

$$\underset{x}{argmax} -D(\theta_s || \theta_x) = \underset{x}{argmax} - \sum_{\omega \in V} p(\omega | \theta_s) \log \frac{p(\omega | \theta_s)}{p(\omega | \theta_x)}$$

where $x \in \{j, P, N\}$, and θ_s is a language model of s

Utilizing the model

- Overall opinion for topic j in a document
 - The overall sentiment distribution for topic j in document d is the sentiment coverage $\{\delta_{j,d,F}, \delta_{j,d,P}, \delta_{j,d,N}\}$
 - The overall sentiment strength for the topic j is,

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

Roadmap

- Introduction
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- **Sentiment Dynamics Analysis**
 - **HMM based approach**
- Experiments and Results
- Summary
- Conclusion

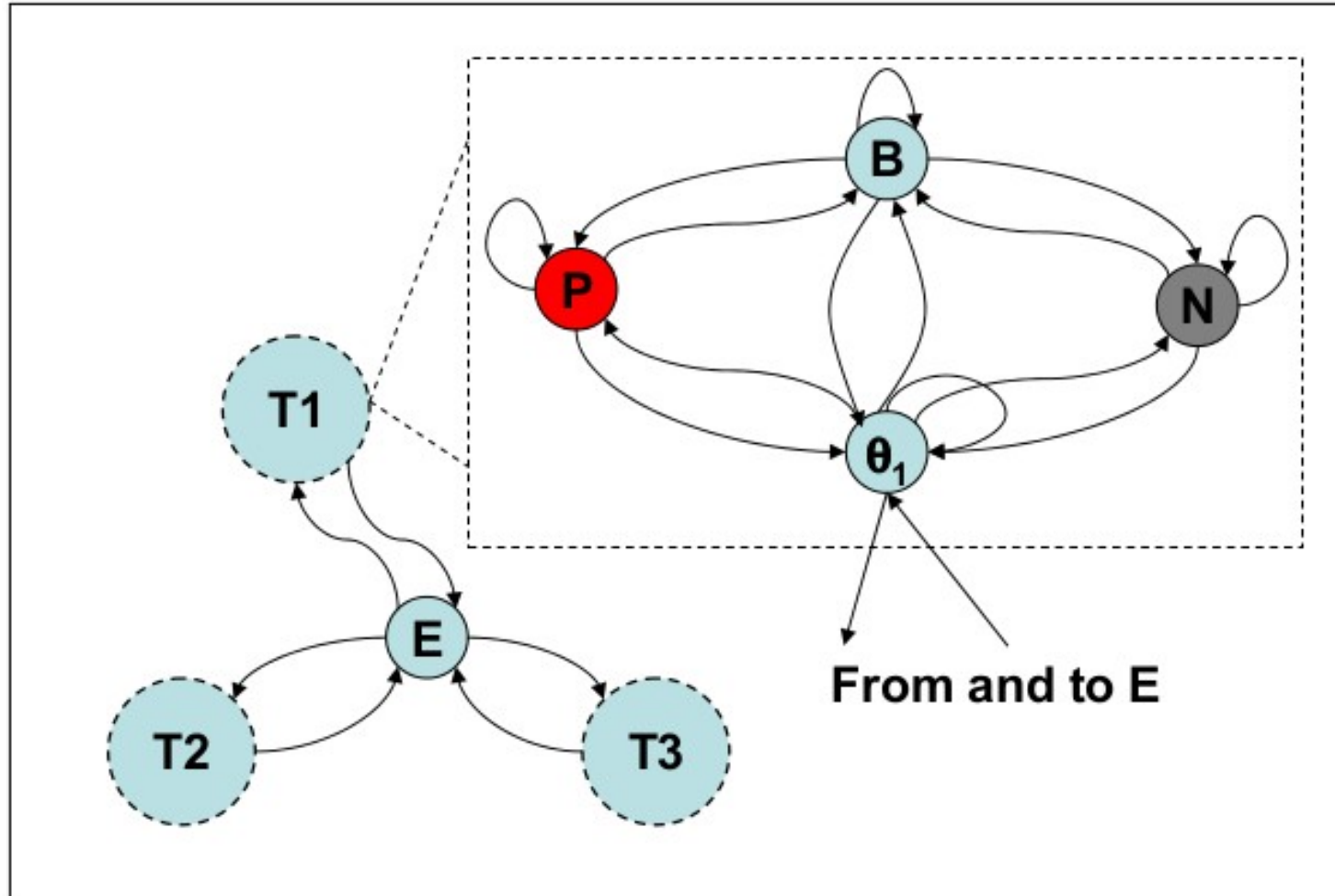
Sentiment Dynamics Analysis

- Sentiment strength over time
- $p(t|\theta_j)$, $p(t|\theta_j, \theta_P)$ and $p(t|\theta_j, \theta_N)$
 - Time period is not there in the original model
- Adding t to the model
 - Will increase free parameters in the model
- HMM based approach is used in this case

HMM based approach

- Sort documents according to time stamp
- Convert whole collection into a sequence of words
- Initial thought
 - Each topic corresponds to a state
 - Output probability of a state j , $p(\omega|\theta_j)$
 - Can add sentiment states
 - Cannot decode which sentiment is about which topic

HMM based approach



Roadmap

- Introduction
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- **Experiments and Results**
 - **Data Sets**
 - **Sentiment Model Extraction**
 - **Topic Model Extraction**
 - **Topic Sentiment Summarization**
 - **Topic Lifecycle and Sentiment Dynamics**
- Summary
- Conclusion

Data Sets

- Two Datasets
 - To learn general sentiment models
 - Opinmind
 - To extract topic models and sentiment dynamics
 - Google Blog Search

Data Sets

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

OPIN Dataset

Data Sets

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

TEST Dataset

Sentiment Model Extraction

- Opinmind search engine is used
 - 10 different topics, 5 queries/topic
 - 10 topic specific datasets
 - 1 Dataset containing mixture of all topics
- OPIN Dataset

Sentiment Model Extraction

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

Sentiment Model Extraction

- Observations
 - Models biased towards specific contents in the topic if dataset contains only one topic
 - 10-topic mixture dataset based model more general than topic specific models

Topic Model Extraction

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

Ipod

Topic Model Extraction

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

Da Vinci Code

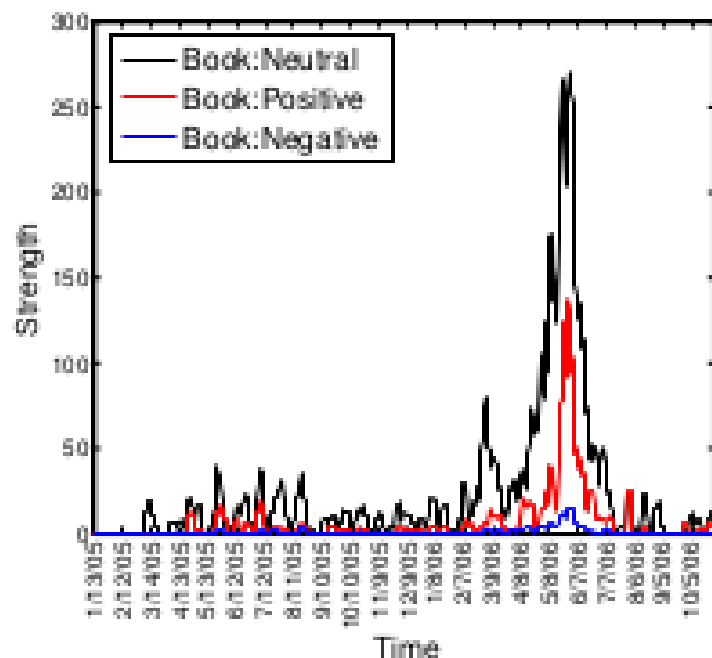
Topic Sentiment Summarization

	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.

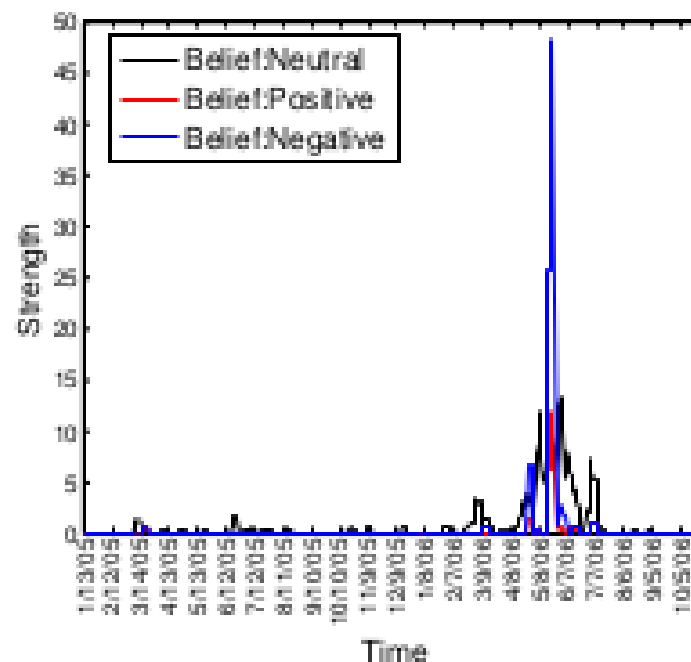
Table 6: Topic-sentiment summarization: Da Vinci Code

TSM			Opinmind	
	Thumbs Up	Thumbs Down	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!!!!	I love my iPod, I love my G5...	I hate ipod.
			I love my little black 60GB iPod	Stupid ipod out of batteries...
2	the battery is one serious example of excellent relibability	Poor battery lifeiPod's battery completely died	I LOVE MY IPOD	" hate ipod " = 489..
			I love my iPod.	my iPod looked uglier...surface...
3	My new VIDEO ipod arrived!!! Oh yeah! New iPod video	fake video ipod Watch video podcasts ...	- I love my iPod.	i hate my ipod.
			... iPod video looks SO awesome	... microsoft ... the iPod sucks

Topic and Sentiment Dynamics

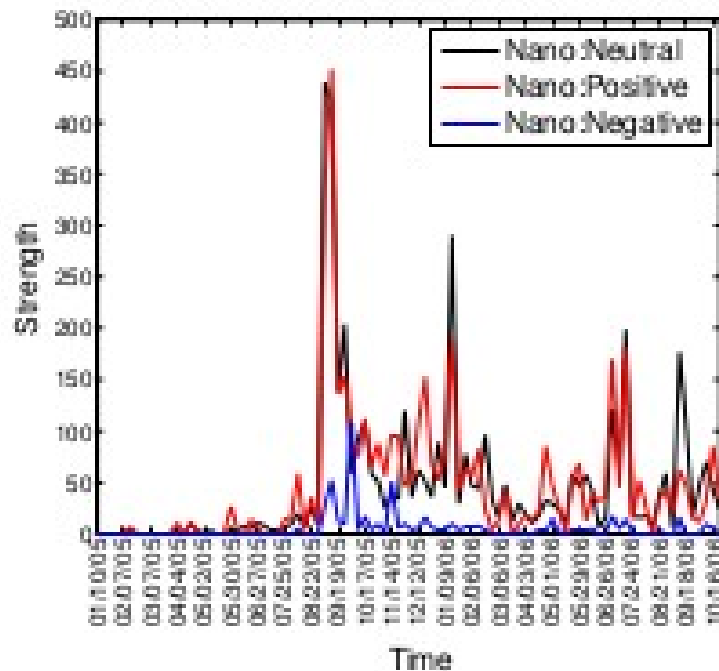


(a) Da Vinci Code: Book

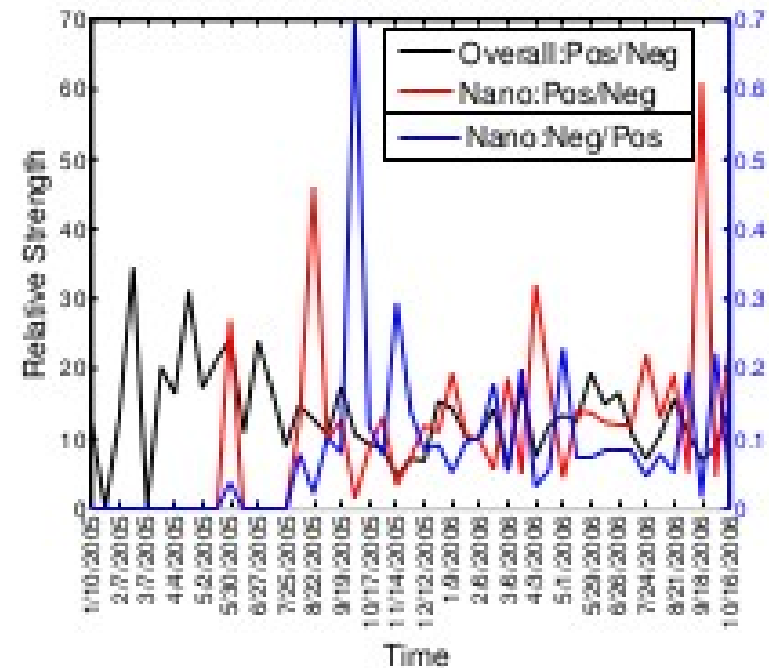


(b) Da Vinci Code: Religion

Topic and Sentiment Dynamics



(c) iPod: Nano



(d) iPod: Relative

Roadmap

- Introduction
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- **Summary**
- Conclusion

Summary

- Formally defined the problem of Topic Sentiment analysis
- Proposed a Topic Sentiment Mixture model to address the problem
- Learnt general Sentiment Models
- Extracted Topic Models with/without priors
- Extracted Topic Lifecycles and associated Sentiment Dynamics

Roadmap

- Introduction
- Motivation
- Problem Formulation
- Topic Sentiment Mixture
- Sentiment Dynamics Analysis
- Experiments and Results
- Summary
- **Conclusion**

Conclusion

- TSM model is effective for Topic Sentiment Analysis
- Generates useful topic-sentiment summaries for blog search
- Possible Extensions
 - Use separate sentiment model for each topic

Thankyou

Other Papers read

- Jo, Yohan, and Alice H. Oh. "Aspect and sentiment unification model for online review analysis." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.
- Lin, Chenghua, and Yulan He. "Joint sentiment/topic model for sentiment analysis." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.
- Zhang, Min, and Xingyao Ye. "A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval." Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.