



# What Users Care about: A Framework for Social Content Alignment

Lei Hou<sup>1</sup>, Juanzi Li<sup>1</sup>, Xiaoli Li<sup>2</sup>, Jiangfeng Qu<sup>1</sup>, Xiaofei Guo<sup>1</sup>, Ou Hui<sup>1</sup>, Jie Tang<sup>1</sup>

<sup>1</sup> *Knowledge Engineering Group, Dept. of Computer Science and Technology, Tsinghua University*

<sup>2</sup> *Institute for Infocomm Research, A\*STAR, Singapore*

# Outline

- Motivation & Challenges
- Related Work
- Approach
- Experiment
- Conclusion & Future Work

# Motivation



78% of Internet users in China (461 million) read news online[Jun, 2013, CNNIC]



The average numbers of comments for top news in Yahoo! and Sina are 5684.6 and 9205.4 respectively (on Nov, 2012)



How to find  
what users  
care  
about



# Motivation

- How to achieve that?
  - Link sentences and comments  $\leftrightarrow$  Social Content Alignment
- How to align?

WASHINGTON—...

Boehner won the backing of 220 Republicans, who retained a majority in the chamber after November's election. But a handful of GOP members **voted no or abstained**. Most Democrats voted for House Minority Leader Nancy Pelosi.

Boehner's grasp on his speakership seemed tenuous going into the **vote**.

....

Several northeastern Republicans loudly criticized Boehner for stalling a **\$60 billion relief bill** for states hit by Superstorm Sandy. Boehner has pledged to hold a vote on Sandy relief on Friday.

....

Once the votes were cast and Boehner was announced the winner, Republican and Democratic leaders joined the Ohio delegation in escorting Boehner to the speaker's chair, where he will **serve for two more years**.

In his first speech to the 113th Congress, Boehner urged members to remain true to the Constitution and focused his remarks on the national **debt**.

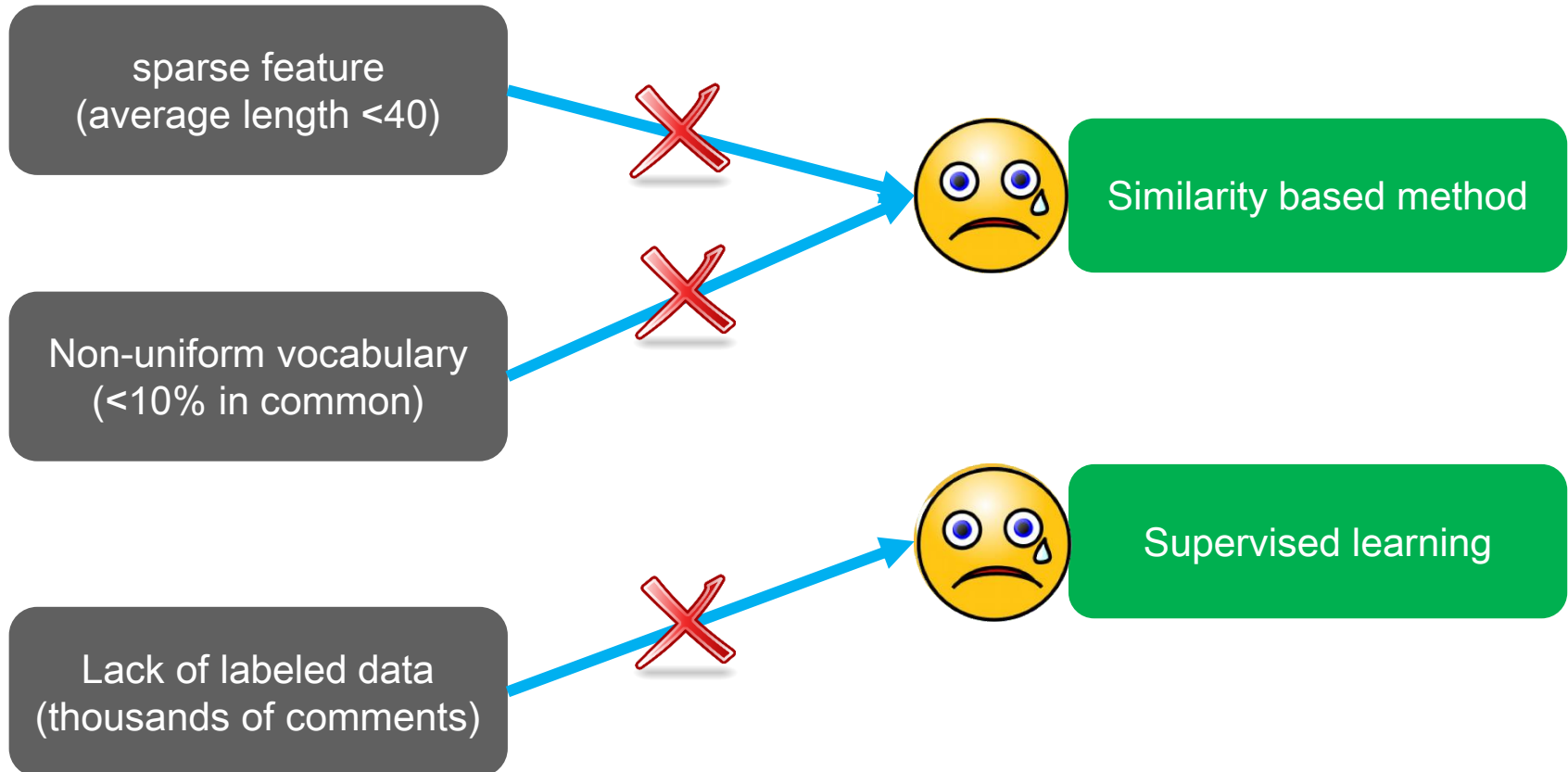
"Our government has built up too much **debt**. Our economy is not producing enough jobs. These are not separate problems," Boehner told the members in the chamber. "At \$16 trillion and rising, our national **debt** is draining free enterprise and weakening the ship of state. "The American Dream is in peril so long as its namesake is weighed down by this anchor of debt. Break its hold, and we begin to set our economy free."

88,055 comments

Popular Now | Newest | Oldest | Most Replied

- How do they include all that outrageous pork in the hurricane relief bill? it's disgusting 22%
- good now stand by your words, no rise in the debt ceiling unless there is major cuts, no pork and no foreign aid. 14%
- CNN is reporting 220 out of 234 voting for Boehner, with 12 declining to vote at all (which is like voting "no") I'm surprised...I would've sworn he would've been voted out, given his party's reaction to the cliff deal... 29%
- The margin was? Yahoo news, worse than MTV news. 26%
- Conservatives demand term limits right up to the moment they are elected. Then "term limits" becomes a dirty word.. Over the next two years they gin up a dozen or so "powerful reasons" why term limits should not apply to them. 9%

# Challenges



# Related Work-social content analysis

- Readalong: reading articles and comments together.
  - Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya.
  - In WWW'11 (poster)
- Supervised matching of comments with news article segments.
  - Dyut Kumar Sil, Srinivasan H. Sengamedu, and Chiranjib Bhattacharyya.
  - In CIKM'11 (short paper)
- Opinion integration through semi-supervised topic modeling.
  - Yue Lu and Chengxiang Zhai.
  - In WWW'08

# Related Work-topic modeling

- A time-dependent topic model for multiple text streams.
  - Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsoulis.
  - In KDD'11
- Multi-topic based query-oriented summarization.
  - Jie Tang, Limin Yao, and Dewei Chen
  - In SDM'09
- Cross-domain collaboration recommendation.
  - Jie Tang, Sen Wu, Jimeng Sun, and Hang Su.
  - In KDD'12,

# Related Work-positive unlabeled learning

- Building text classifiers using positive and unlabeled examples.
  - Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu.
  - In ICDM'03
- Learning with positive and unlabeled examples using weighted logistic regression.
  - Wee Sun Lee and Bing Liu.
  - In ICML'03.
- Learning to classify texts using positive and unlabeled data.
  - Xiaoli Li and Bing Liu.
  - In IJCAI'03.
- Learning to identify unexpected instances in the test set.
  - Xiaoli Li, Bing Liu, and See-Kiong Ng.
  - In IJCAI'07.



# Approach Framework

## PHASE 1

Document  
Comment  
Topic Model



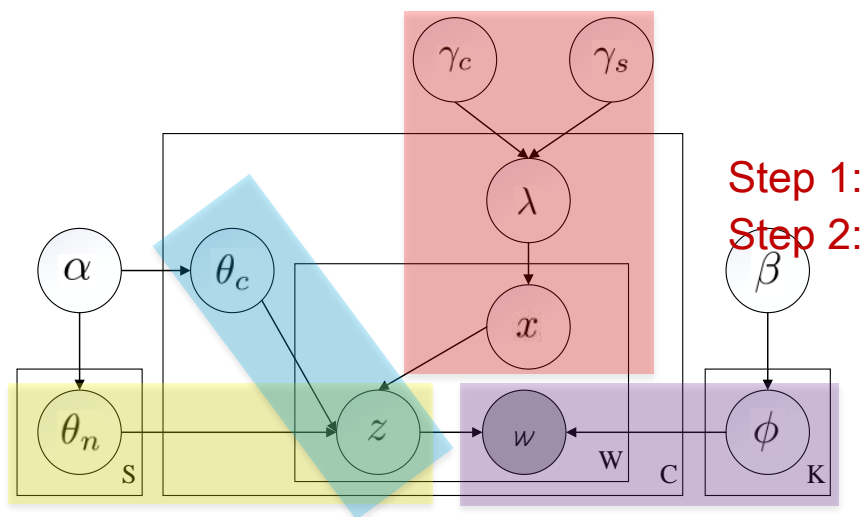
## PHASE 2

Learning from  
Positive and  
Unlabeled Data

- Different vocabulary
- Sparse feature
- Dependency

- Unbalanced volume
- Lack of labeled data

# Document-Comment Topic Model



Top words for topic *launch cost*

<div></div> <div></div> <div></div> <div></div> <div></div>	Aid	<div></div> <div></div> <div></div> <div></div> <div></div>
	Stomach	
	America	
	Food	
	Korea	
	Korea	<div></div> <div></div> <div></div> <div></div> <div></div>
	Money	
	Launch	
	America	
	Food	

- Comment only
- News only
- Both

The left only uses comments, and the right takes news as background

**Algorithm 1:** Generative process for DCT model.

**Input:** the priors  $\alpha, \beta, \gamma_c, \gamma_s$ ;  $S$  and  $C$

**Output:** estimated parameters  $\theta_s, \theta_c, \lambda$  and  $\phi$

Initialize a standard LDA model over  $S$ ;

```

foreach comment  $c \in C$  do
  foreach word  $w_{ci} \in c$  do
    Toss a coin  $x_{ci}$  according to
     $\text{bernoulli}(x_{ci}) \sim \text{beta}(\gamma_s, \gamma_c)$ , where  $\text{beta}(\cdot)$  is
    a beta distribution, and  $\gamma_c$  and  $\gamma_s$  are two
    parameters;
    if  $x_{ci} = 0$  then
      Draw a topic  $z_{ci} \sim \text{multi}(\theta_c)$  from a
      comment-specific topic mixture;
    else
      Draw a topic  $z_{ci} \sim \text{multi}(\theta_s)$  from a
      document-related topic mixture;
    end
    Draw a word  $w_{ci} \sim \text{multi}(\phi_{z_{ci}})$  from
     $z_{ci}$ -specific word distribution;
  end
end
  
```

# PU Learning

---

## Algorithm 2: PU learning

---

**Input:** news sentences  $S$ , social contents  $C$ , topic distribution  $\theta$ , word distribution  $\phi$

**Output:** A set of classifiers

**for each topic do**

1. Extract the positive and unlabeled example set;

2. Build first classifier:

- calculate centroid and radius to construct a hyper-sphere
- extract potential positive examples and negative examples
- build first classifier using *Ricchio*

3. Build final classifier:

- classify unlabeled data using first classifier
- build final classifier using WSVM

**end**

---

s & c \ topic	vote	relief	...	debt
$S_1$	<b>0.173</b>	0.039	...	0.094
$S_2$	0.082	<b>0.127</b>	...	0.077
...				
$S_M$	<b>0.184</b>	0.083	...	0.105
$C_1$	...	...	...	...
$C_2$	...	...	...	...
...				
$C_N$	...	...	...	...

## Positive example for topic *vote*

1. But a handful of GOP members voted no or abstained.
2. Boehner's ... seemed tenuous going into the vote.
3. Once the votes were cast and ... .
- ...

# PU Learning

---

## Algorithm 2: PU learning

---

**Input:** news sentences  $S$ , social contents  $C$ , topic distribution  $\theta$ , word distribution  $\phi$

**Output:** A set of classifiers

**for each topic do**

1. Extract the positive and unlabeled example set;

2. Build first classifier:

- calculate centroid and radius to construct a hyper-sphere
- extract potential positive examples and negative examples
- build first classifier using *Ricchio*

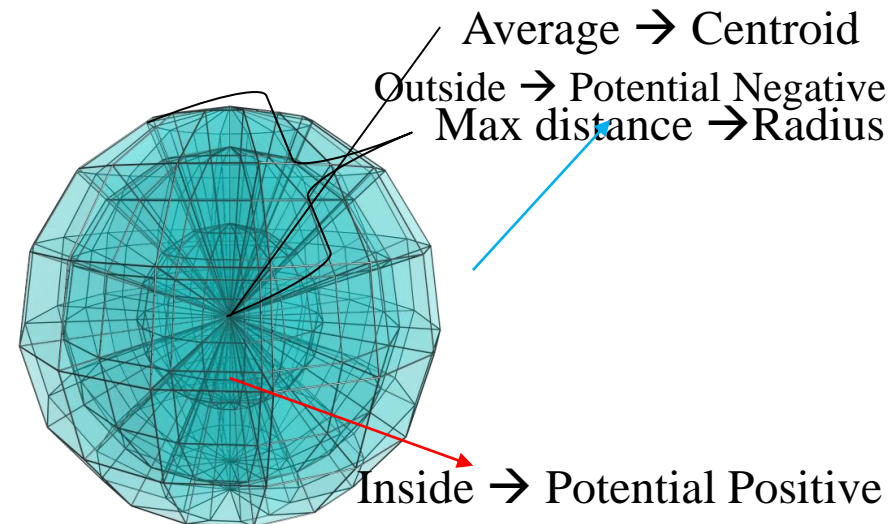
3. Build final classifier:

- classify unlabeled data using first classifier
- build final classifier using WSVM

**end**

---

	$f_1$	$f_2$	...	$f_K$
$P_1$	0.043	0.019	...	0.024
$P_2$	0.052	0.037	...	0.017
...				
$P_{ P }$	0.054	0.033	...	0.015



# PU Learning

---

**Algorithm 2:** PU learning

---

**Input:** news sentences  $S$ , social contents  $C$ , topic distribution  $\theta$ , word distribution  $\phi$

**Output:** A set of classifiers

**for each topic do**

1. Extract the positive and unlabeled example set;
2. Build first classifier:

- calculate centroid and radius to construct a hyper-sphere
- extract potential positive examples and negative examples

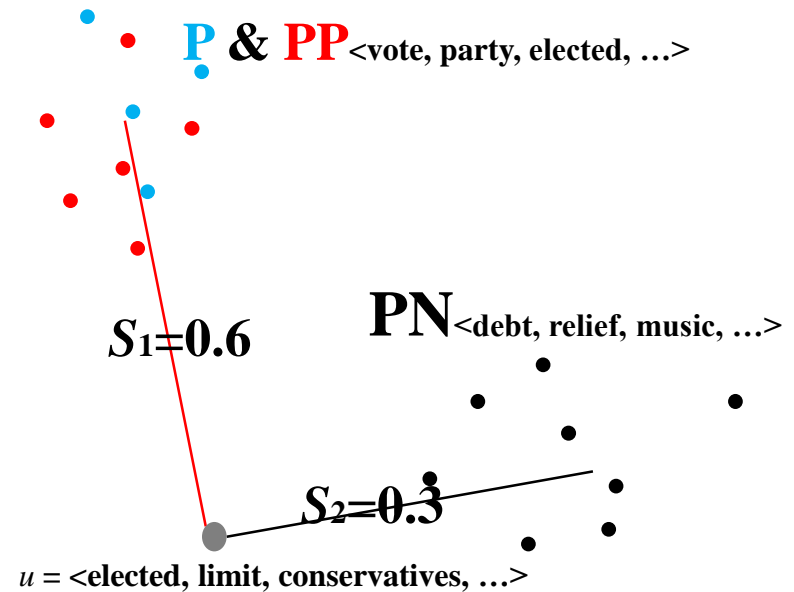
- build first classifier using *Ricchio*

3. Build final classifier:

- classify unlabeled data using first classifier
- build final classifier using WSVM

**end**

---



Adjust the label according to  $S_1$  and  $S_2$ , as well as assign a confidence score

$$L = \frac{\max(s_1, s_2)}{s_1 + s_2}$$

# PU Learning

---

## Algorithm 2: PU learning

---

**Input:** news sentences  $S$ , social contents  $C$ , topic distribution  $\theta$ , word distribution  $\phi$

**Output:** A set of classifiers

**for each topic do**

1. Extract the positive and unlabeled example set;
2. Build first classifier:

- calculate centroid and radius to construct a hyper-sphere
- extract potential positive examples and negative examples

- build first classifier using *Ricchio*

3. Build final classifier:

- classify unlabeled data using first classifier
- build final classifier using WSVM

**end**

---

	$\mathbf{L}$	$f_1$	$f_2$	$\dots$	$f_K$
$P_1$	1	0.043	0.019	$\dots$	0.024
$P_2$	1	0.052	0.037	$\dots$	0.017
$\dots$					
$LP_1$	0.7	0.054	0.033	$\dots$	0.015
$\dots$					
$LN_1$	0.83	0.003	0.061	$\dots$	0.055
$\dots$					

$$\begin{aligned}
 \text{Minimize : } & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C_P \sum_{i \in P} \xi_i + \\
 & C_{LP} \sum_{j \in LP} \xi_j + C_{LN} \sum_{k \in LN} \xi_k \\
 \text{subject to : } & y_i (\mathbf{w}^T \vec{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, n
 \end{aligned}$$

# Data Set

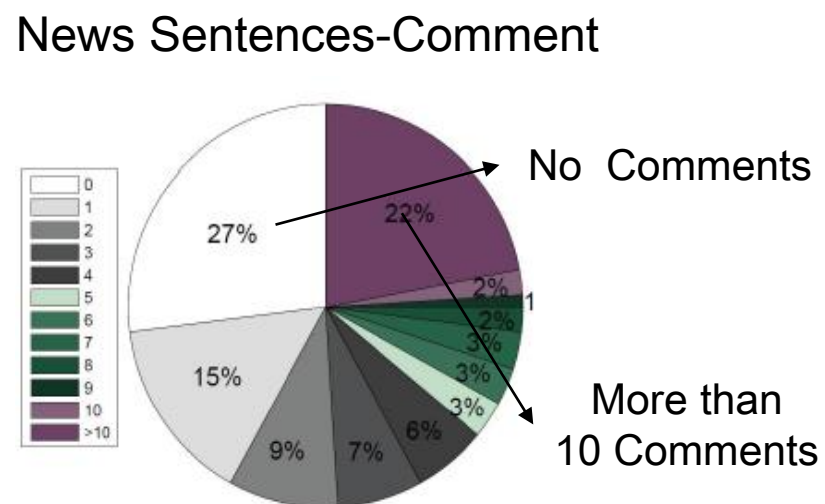
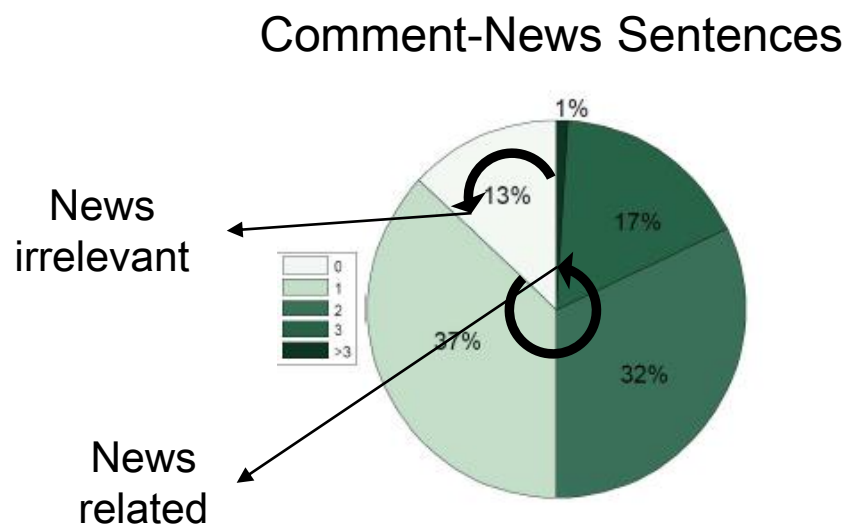
- Sources (Chinese: Sina, English: Yahoo!)
- 22 news articles (10 Chinese, 12 English)
- 950 news sentences (516 in Chinese, 434 in English)
- 6,219 comments (4,069 in Chinese, 2,150 in English)

Table 1: Statistics on datasets

Source		#Sen/Com	Words	Vocabulary
Sina	<i>Sen</i>	516	8,932	2,772
	<i>Com</i>	4,069	112,853	13,891
Yahoo!	<i>Sen</i>	434	5,767	2,679
	<i>Com</i>	2,150	39,917	9,972

# Annotation

- Manually Annotation
  - 7 annotators (publish task online)
  - Confidence: 5 out of 7 agree
  - Results: 7,520 (cn) + 2,327 (en) links
- Annotated Data Observation





# Baseline Methods & Metric

- Methods

- unsupervised

- VSM: tf-idf + cosine similarity
    - DCT: topic directly

- supervised

- BSVM: classifier on sentence
    - T-SVM: classifier on topic

- Ours(T-PU): unsupervised classifier on topic

- Metric

$$Precision = \frac{|\bigcup_{i=1}^N \{c_i | r_i \cap \tilde{r}_i \neq \emptyset\}|}{|C|}$$

where  $r_i$  and  $\tilde{r}_i$  stands for the annotated alignments and the alignments that found by our method

# Results

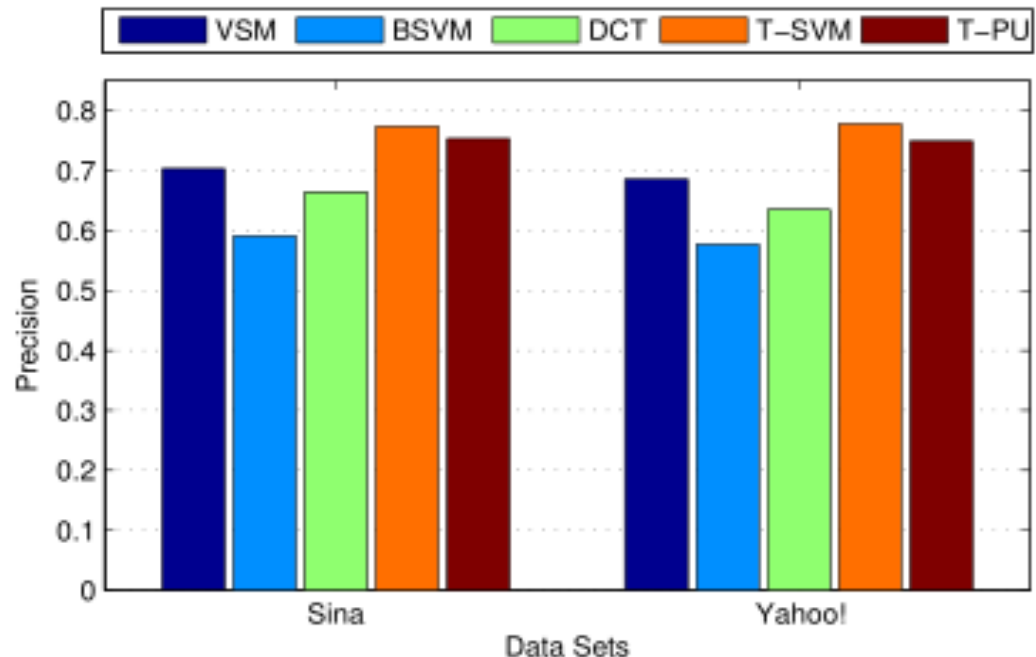
- Overall

Table 2: Overall results on two datasets

	Precision	Recall	F1-Measure
Sina	75.3%	56.7%	64.7%
Yahoo!	74.9%	63.4%	68.7%

- Comparison

- best among unsupervised methods (VSM +7.9%)
- BSVM (+25.9%), significant improvement
- T-SVM, comparable results (-2.1% in Sina and -2.9% in Yahoo!)



# Results

- What leads to failed alignment
  - comment chain (a series of comments issued by two or more users while discussion)
  - topic drift
- Example:

**boner** 1 day ago 151 ▲ 6 ▼  
A cold war in a new era (even more so I guess than it already is). It'll be a new frontier given the power the internet and other technologies the people have at their disposal.

I am sure if we head down that road, the people of Russia and the people of the USA will see through their respective governments BeeEss tactics and unite...the people are more powerful.

[More](#)  
[Collapse Replies \(13\)](#) [Reply](#)

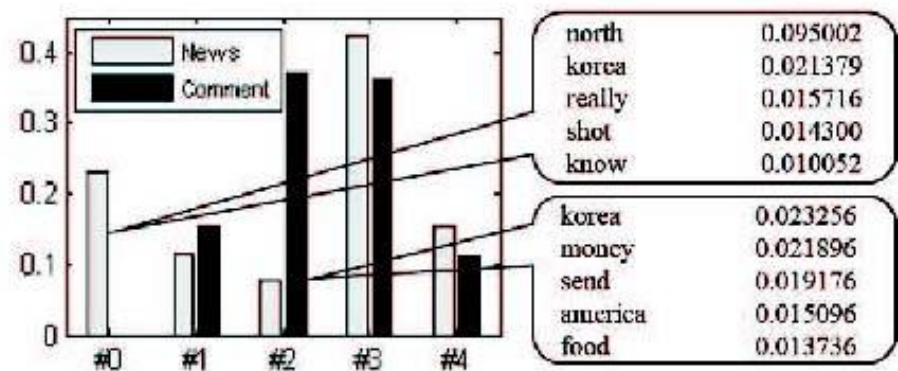
**Van K** 21 hours ago 6 ▲ 5 ▼  
Sounds ideal. In reality, the people in this country have not prevented their government from enacting the FISA and Patriot Acts. The people in Russia did not prevent their government from enacting bans on speech like imprisoning Punani Riot and outlawing support of gay rights--nor in preventing far more surveillance on their citizens than goes on in the U.S.

[More](#)

**AD123** 18 hours ago 5 ▲ 2 ▼  
We need for form a website or some sort of gathering place FAST. They WILL be on the lookout for these types of comments VERY SOON as they do not have much time before heads will snap. PLEASE WE MUST ACT FAST.

**Jason** 14 hours ago 4 ▲ 0 ▼  
I was reading your post so seriously and then I read your name and burst out laughing

**Elizabethanne** 12 hours ago 8 ▲ 1 ▼  
I hope you are right. Americans seem to behave like sheeple as of late.



# Conclusion

- Study the social content alignment problem and present a two-phase framework to address it
- Propose DCT model which exploits Web document, social content and their dependency
- Employ PU learning algorithm for alignment
- Experimental results show the effectiveness of the proposed approach

# Future Work

- Alignment over similar web documents
- Whether the social relationships influence the alignment
- Topic drift in the social content

**Thanks!**