Artificial Intelligence in Automated Authoring & Health Informatics

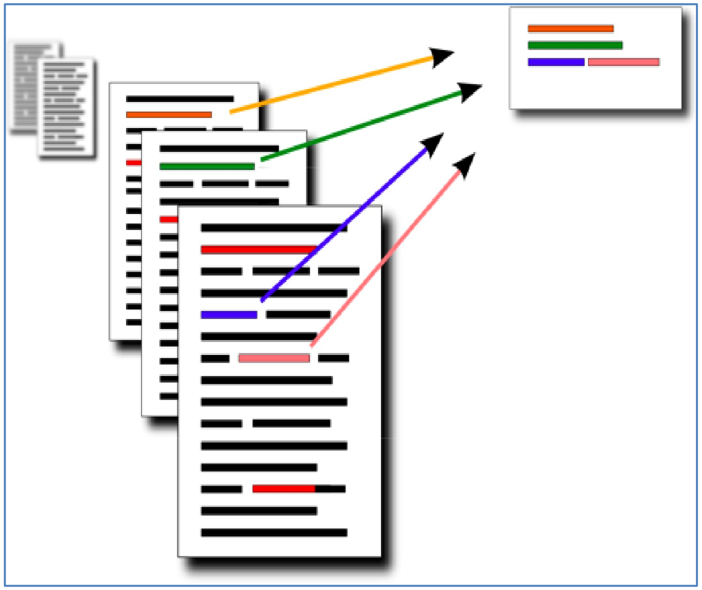
Prasenjit Mitra

Professor

College of IST

Research Questions

- How can we summarize (abstractive) content from any domain?
  - *News, Conversations, Social media, etc.*
- Can we use abstractive summarization to automatically author content?
  - *For example, Wikipedia articles.*



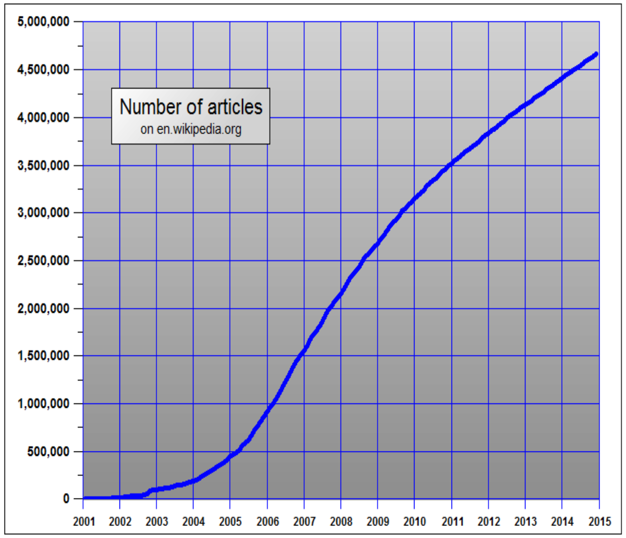
Motivation: Extractive vs Abstractive

- **Extractive**
  - *"Extracting" certain sentences from the original input*
- **Abstractive**
  - *Involves text understanding, rewriting in own words*
  - *More information (Genest and Lapalme, 2011)*

Application: Wikipedia Article Authoring

Why is it important?

- **5 million** articles in the English Wikipedia



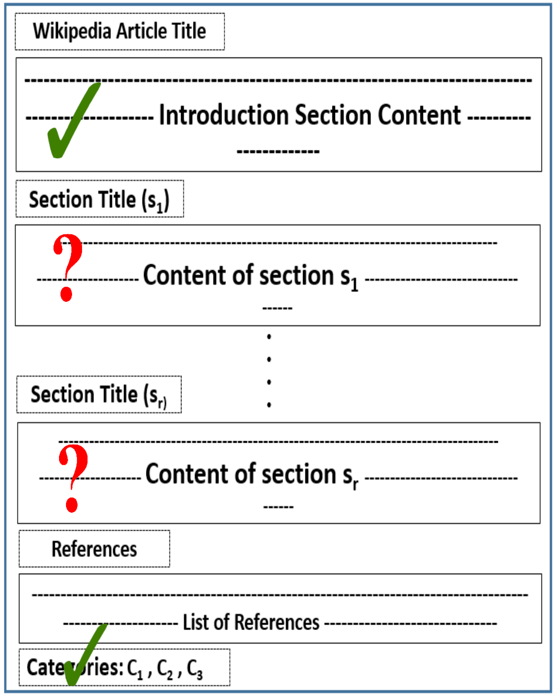
Around 700 articles being added per day!

- rticles keep growing: **Backlog!**

Train computers to fill in the information gap!

Identifying Templates

Improving stubs



Learning from Wikipedia Categories
Text Classification – Web content to sections
Summarization

Creating new articles

The ISU Representative was Olaf Poulsen and the ISU Technical Delegate was Josef Dědič.

The ISU representative was John R. Shoemaker (USA) and the ISU Technical Delegate was Elemer Tertak (Hungary).

Fabio Bianchetti

Fabio Bianchetti is a member of the International Skating Union Tech
2006 Winter Olympics Opening Ceremony

He is the son of Sonia Bianchetti, also a long-time former ISU official.

Similar entities are mentioned in similar contexts

Distributional semantics
Red-linked entity: Sonia Bianchetti
Existing: Josef Dedic and Elemer Tertak

Paragraph2Vec – Le and Mikolov' 2014
- Represents words (entities) as vectors
- Can also represent long sequences

Content Assignment and Generation

Content from web

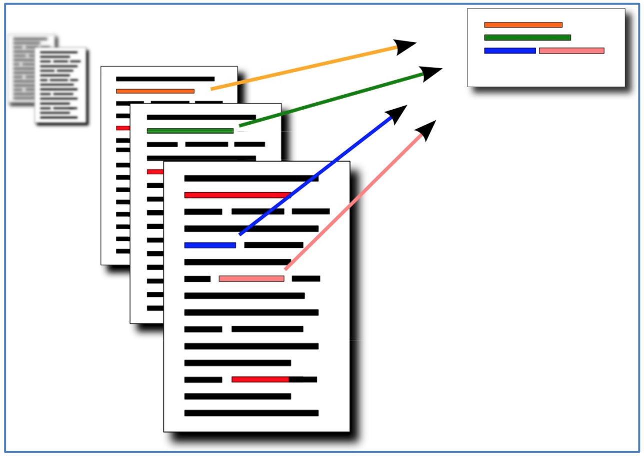
- Query Reformulation
- Cleaning (Boilerplate detection)
- **Text Classification**
  - Stubs (WikiKreator, ACL' 2015):
    - ✓ Topic model distribution vectors as features
  - New Articles (WikiWrite, IJCAI' 2016):
    - ✓ Paragraph Vectors as features

Summarization (Abstractive)

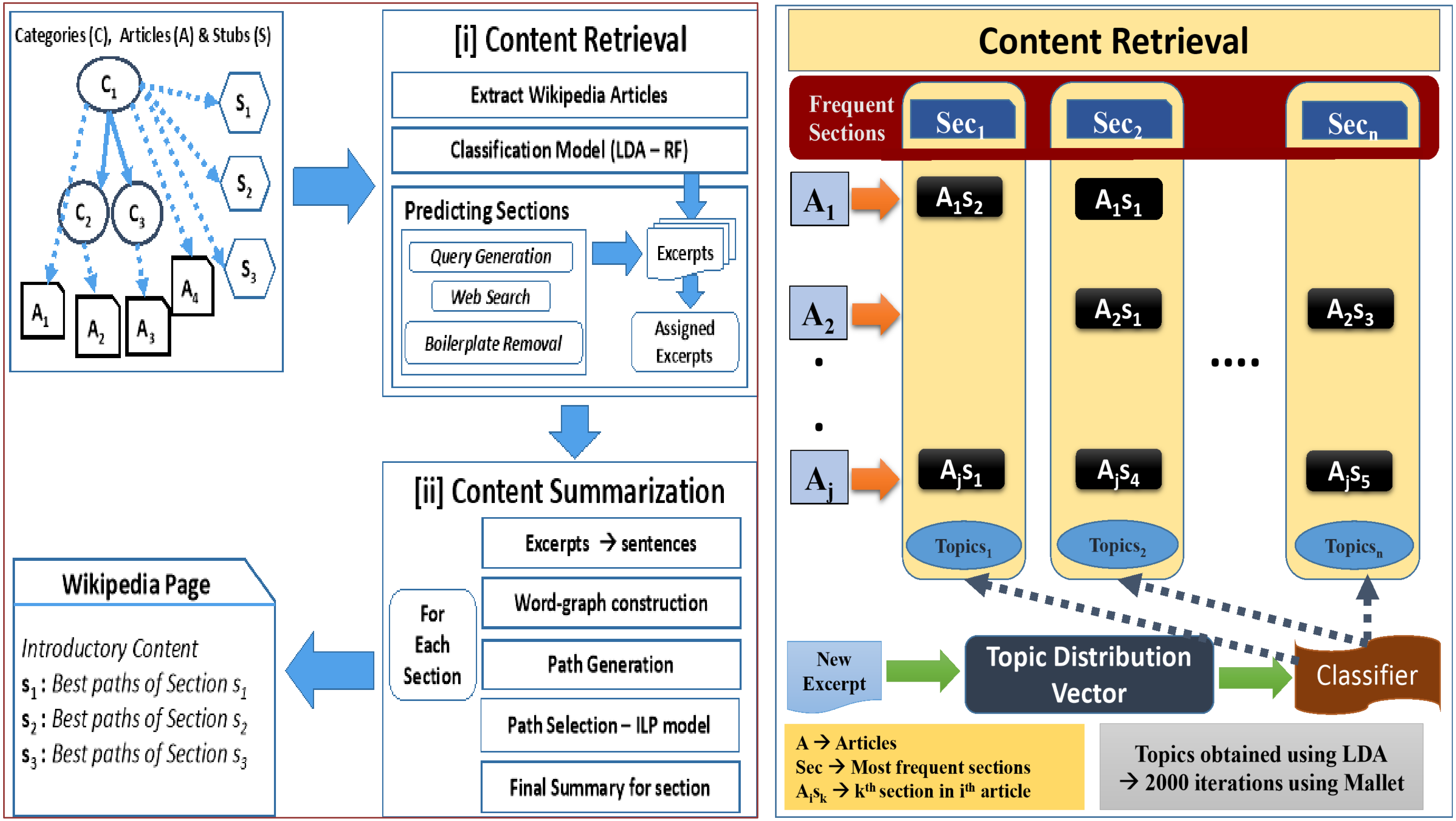
✓ **Copyright Violation [Banerjee et. al, 2014]**

✓ Combine information from multiple documents

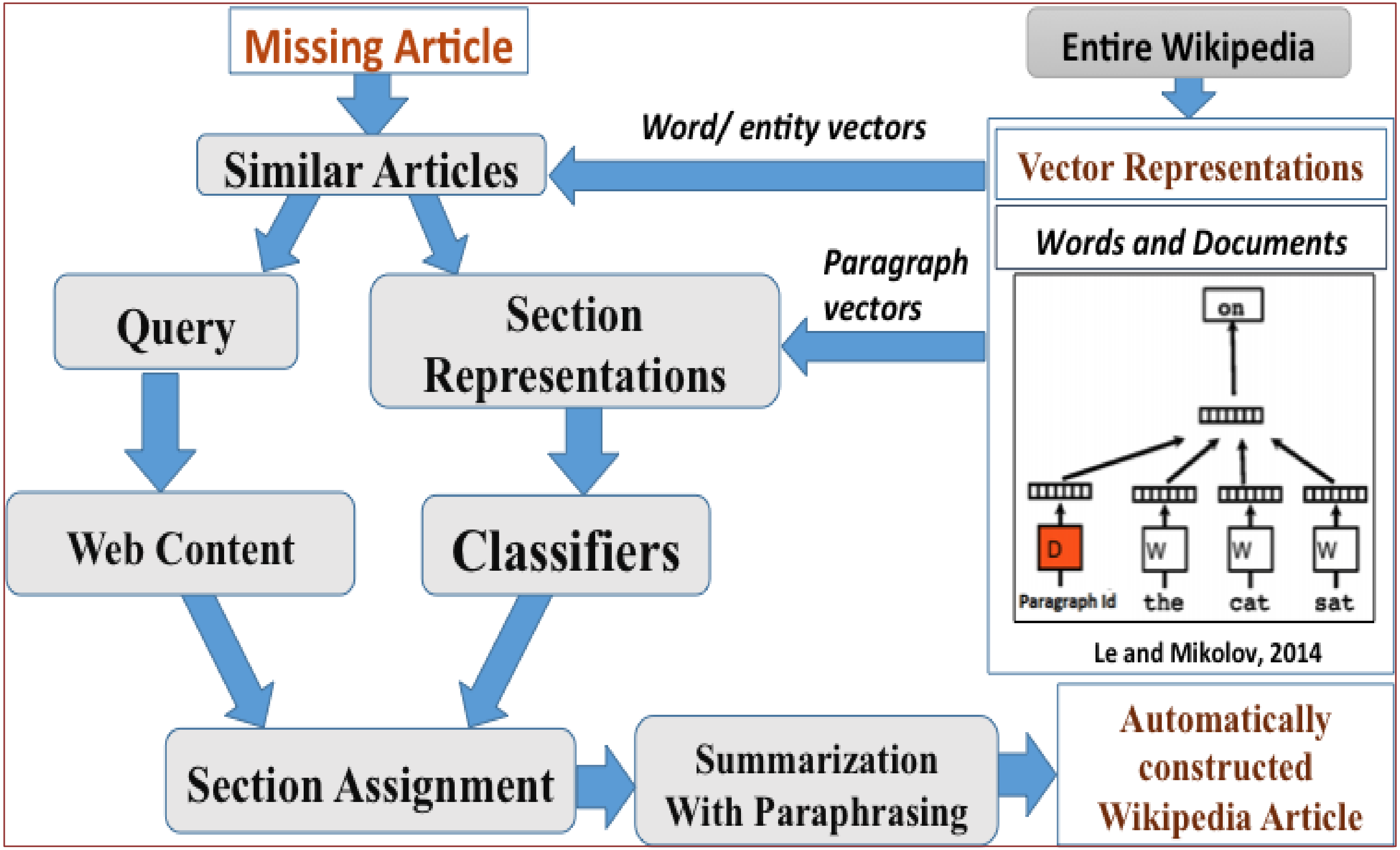
✓ Appear human-like



WikiKreator Framework (Banerjee and Mitra, 2015)



WikiWrite Framework (Banerjee and Mitra, 2016)



Results – Wikipedia Section Classification

Category	Most Frequent Sections
American Mathematicians	*Awards, Awards and honors, Biography, Books, Career, Education, Life, Publications, Selected publications, Work*
Diseases and Disorders	*Causes, Diagnosis, Early life, Epidemiology, History, Pathophysiology, Prognosis, Signs and symptoms, Symptoms, Treatment*
US Software companies	*Awards, Criticism, Features, Games, History, Overview, Products, Reception, Services, Technology*

F1-scores

Category	LDA-RF	SVM-WV
American Mathematicians	**0.778**	0.478
Diseases and Disorders	**0.886**	0.801
US Software companies	**0.880**	0.537

Results – Wikipedia Article Generation

- WikiKreator *(System to improve stubs)*

Category	System	ROUGE-1	ROUGE-2
American Mathematicians	WikiKreator	**0.522**	**0.311**
	Perceptron	0.431	0.193
	Extractive	0.471	0.254
Diseases and Disorders	WikiKreator	**0.537**	**0.323**
	Perceptron	0.411	0.197
	Extractive	0.473	0.232
US Software companies	WikiKreator	**0.521**	**0.321**
	Perceptron	0.421	0.228
	Extractive	0.484	0.257

Statistics	Count
Number of stubs edited	40
Number of stubs retained without any changes	21
Number of stubs that required minor editing	6
Number of stubs where edits were modified by reviewers	4
Number of stubs in which content was removed	9
Average change in size of stubs	515 bytes
Average number of edits made post content-addition	~3

Table 4: Statistics of Wikipedia article generation

Statistics	
Number of articles in mainspace	47
Entire edit retained	12
Modification of content	35
Average number of edits	11
Percentage of references retained	72%

Appended stubs with paraphrasing

Statistics	WikiKreator	WikiWrite
Number of stubs appended	40	40
Entire edit retained	15	32
Modification of content	5	5
Content Removed	20	3
Avg. change in size	287 bytes	424 bytes
Avg. no of edits	3.82	1.39

Automatic authoring can be useful in several other domains.
Example: Automatic news articles for journalists to create a final article based on a stub.

Example – Stub Improvement

Actinic conjunctivitis

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Cdale0112 (talk | contribs) at 03:27, 22 May 2014 *(Added image to page)*. The present address (URL) is a pe
may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

Actinic conjunctivitis is an inflammation of the eye contracted from prolonged exposure to actinic (ultraviolet) rays. Symptoms are redness and swelling of the eyes. Most often the condition is caused by prolonged exposure to Klieg lights, therapeutic lamps, or acetylene torches. Other names for the condition include Klieg conjunctivitis, eyeburn, arc-flash, welder's conjunctivitis, flash keratoconjunctivitis, actinic ray ophthalmia, x-ray ophthalmia, and ultraviolet ray ophtha

References

1. ^ "Dorland's Medical Dictionary (confabulation - connexus)". Archived from the original

See also

- Conjunctivitis
- Photokeratitis

Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Español
Hrvatski
Edit links

acetylene torches. Other names for the condition include Klieg conjunctivitis, eyeburn, arc-flash, welder's conjunctivitis, flash keratoconjunctivitis, actinic ray ophthalmia, x-ray ophthalmia, and ultraviolet ray ophthalmia.[1]

Contents [hide]

Causes

Conjunctivitis is prevalent among children of the highlands of Ecuador. The finding supports the hypothesis that exposure is the main cause of the disease. [2]

Symptoms

Conjunctivitis eye condition contracted from exposure to Actinic rays. Symptoms are redness and swelling. [3]

References

1. ^ "Dorland's Medical Dictionary (confabulation - connexus)". Archived from the original on 13 August 2007. Retrieved 2007-07-27.
2. ^ http://www.ncbi.nlm.nih.gov/pubmed/19393514
3. ^ http://medicine.academic.ru/113418/actinic_conjunctivitis

See also

- Conjunctivitis
- Photokeratitis

live-2433-3730-jair.ps

BDmusic23.Com_Be....mkv
Canceled

A starting point for authors!

Example – New Article (Atripliceae)

Atripliceae

From Wikipedia, the free encyclopedia

Atripliceae are a tribe of the subfamily Chenopodioideae belonging to the plant family Amaranthaceae. *Atriplex* is the largest genus of the family. Species of Atripiceae are ecologically important in steppe and semi-desert climates.[1]

Distribution [edit]

Most of the species are distributed in Africa, Australia, and North America, with some others spread out worldwide.[2]

Taxonomy [edit]

Traditional taxonomy of Atripliceae based on morphological features has been controversial.[1][2] Molecular studies have found that many genera are not true clades. One such study found that Atripliceae could be divided into two main clades, Archiatriplex, with a few, scattered species, and the larger Atriaplex clade, which is highly diverse and found around the world.[2]

References [edit]

1. ^ a b "Molecular phylogeny of Atripliceae (Chenopodioideae, Chenopodiaceae): Implications for systematics, biogeography, flower and fruit evolution, and the origin of C4 photosynthesis.". *www.pubfacts.com*. Retrieved 2015-12-30.
2. ^ a b c Flores, Hilda; Davis, Jerrold I. "A Cladistic Analysis of Atripliceae (Chenopodiaceae) Based on Morphological Data". *Journal of the Torrey Botanical Society*. **128** (3). doi:10.2307/3088719.

This is an improved version after several edits…

Deep Learning for Wikipedia Generation (ongoing)

Frederick George Emmison was born in Bedford on 28 May 1907.[1] He was educated at Bedford Modern School where he excelled academically but was forced to abandon hopes of a University education when his father mistakenly thought a family investment had failed.[4][5]

- ❑ **Given sources 1, 4 and 5, can we reconstruct this paragraph?**
- ❑ **Crawl information from the websites**
  - ❑ Find very similar sentences using cosine similarities.

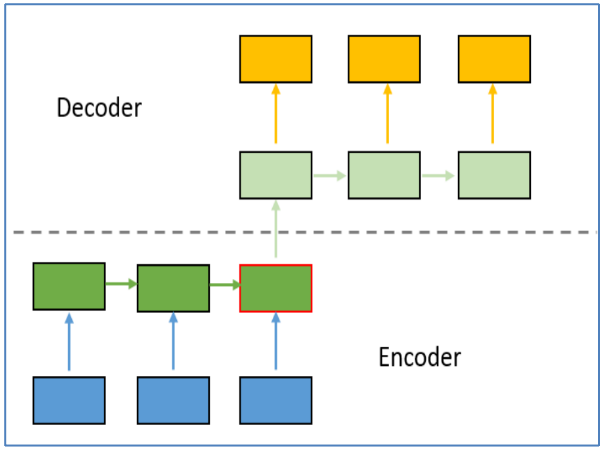
Formulate a Sequence-to-Sequence Problem!
But with a twist.

Multi-source: ***One sentence in Wikipedia can be a result of multiple sentences***

Encoded vectors of multiple sources are decoded to generate the final sentence.

Multi-sequence to Single-Sequence (ongoing)

Sequence to Sequence (Sutskevar et. al, 2014)

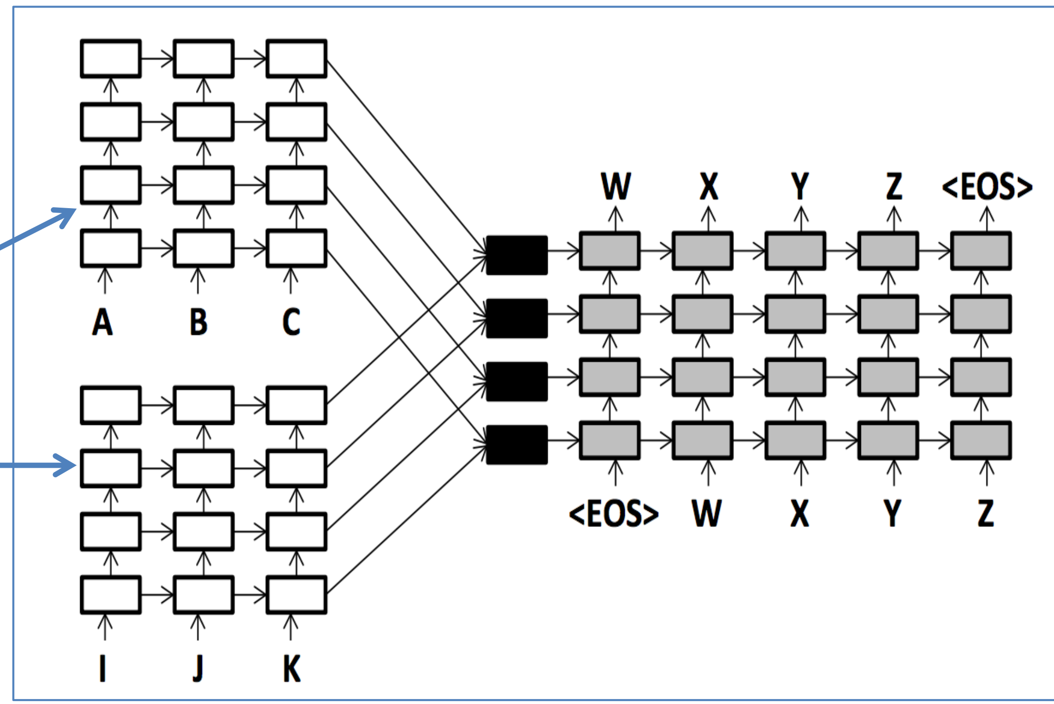


- ❑ Experimenting with LSTM and attention models
- ❑ Trying to replace <UNK> words in output

Multi-Sequence to single-sequence

Sentence 1

Sentence 2



Implementation using *Keras and Theano*

Abstractive Summarization

Contributions: Abstractive Summarization

- ❑ **Proposed Techniques for task-specific summarization**
  - ✓ **News document summarization**
    - ❑ Banerjee et. al, 2015 (IJCAI' 2015)
    - ❑ Cluster similar sentences
    - ❑ ILP formulation maximizing informativeness and readability
  - ✓ **Abstractive Meeting summarization**
    - ❑ Banerjee et. al, 2015 (WWW' 2015, DocEng' 2015)
    - ❑ Dependency-edges based fusion
  - ✓ **Disaster-event Twitter Summarization**
    - ❑ Rudra et. al, 2016 (ACM HyperText' 2016)
    - ❑ Ranking and optimizing content
    - ❑ Focusing on important entities

Framework is easy to adapt across domains – bits and pieces can be modified.

Abstractive Summarization Evaluation

Automatic: ROUGE – Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{Referemce\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Manual: Ask human judges and rate summaries on quality

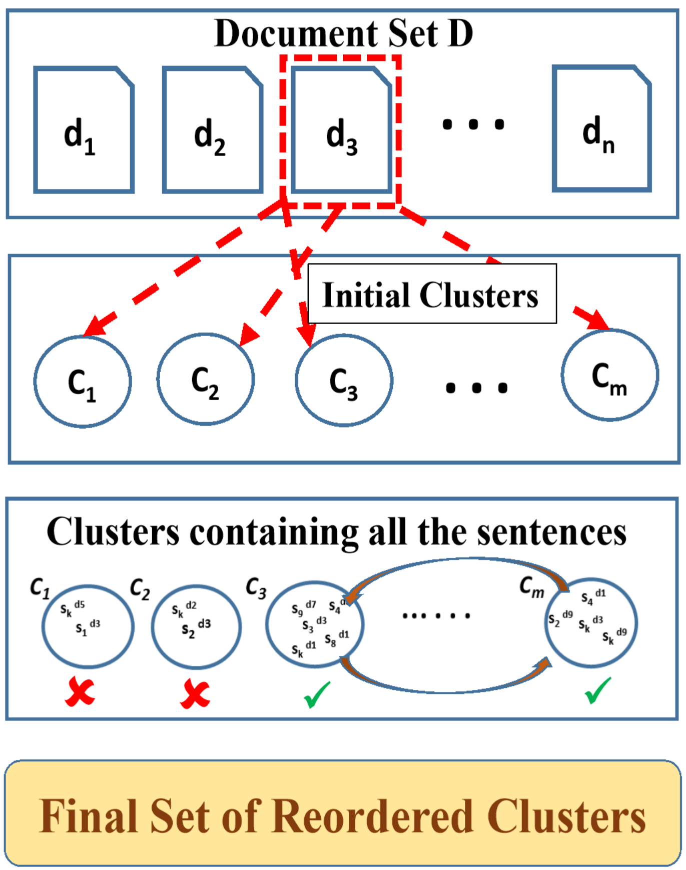
- Informational, Linguistic, etc.

Experiments with three different domains

- **News summarization**
  - ✓ DUC 2004/ 2005 datasets: Outperformed existing systems (ROUGE scores)
  - ✓ Acceptable linguistic quality (Manual evaluation)
- **Meeting summarization**
  - ✓ AMI Corpus: Outperformed other data-driven abstractive system and baseline extractive system (ROUGE scores)
  - ✓ Acceptable linguistic quality (Manual evaluation)
  - ✓ Lots of room for improvement in quality
- **Twitter summarization of disaster specific tweets**
  - ✓ Better ROUGE scores than all other existing methods on disaster-tweet summarization
  - ✓ Manual evaluation using CrowdFlower showed major improvements over other methods: Redundancy, Information, Readability

News-document summarization

Step 1: Clustering



Step 1: Find the most important document $\mathbf{D_{imp}}$ in **D**

Three proposed Methods:

✓ **LexRank** [Erkan and Radev, 2004], **Maximum Pairwise Cosine Similarity** and **Overall Document Collection Similarity**

Step 2: Initialize each sentence in $\mathbf{D_{imp}}$ to different clusters

✓ Each cluster now contains one sentence

Step 3: Align sentences from other documents

✓ Sentences assigned to the cluster based on cosine similarity

Step 4: Prune and reorder clusters

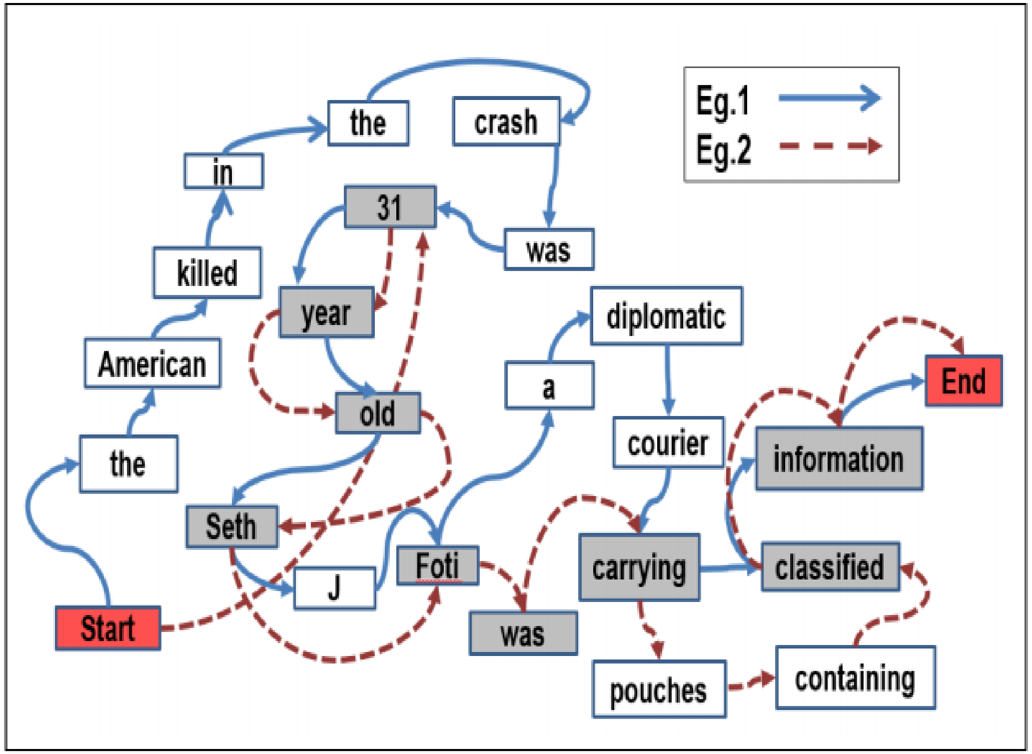
✓ Remove clusters that have less than |D|/2 sentences

✓ Reordering – Majority ordering and Average Position Ordering

Fusion-based New sentence generation

- Construct new sentences using all sentences assigned to a section

1. ***The American killed in the crash was 31-year-old Seth J. Foti**, a diplomatic courier carrying classified information.*
2. *31-year-old Seth **Foti was carrying pouches containing classified information.***



Graph Construction

- Nodes are words (along with POS)
- Edges are adjacencies
- Traverse graph from start to end
- We improved word-graph based fusion using contextual similarity

Select and order sentences

- Informativeness
- Linguistic Quality
- Coherence

The American killed in the crash was 31-year-old Seth J. Foti was carrying pouches containing classified information.

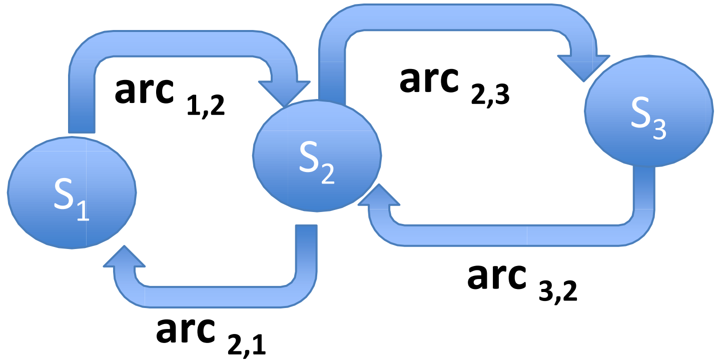
Optimizing summarization quality (Banerjee and Mitra, 2016)

Paths created from the graph – p_i

Multiple factors:

- ✓ I – Informativeness (TextRank, 2004)
- ✓ Sim^{intra} – Pairwise similarity
- ✓ LQ – Language model score

Integer Linear Programming to the rescue!



Regression to assign cosine scores – random sentences as 0 and actual adjacent sentences as 1

Maximize

$$F = \sum_{i=1}^{K} w^{p_i} \cdot p_i + \lambda \sum_{a_{i,j} \in A} coh_{i,j} \cdot arc_{i,j}$$

$$w^{p_i} = I^{p_i} \cdot Sim^{intra}(p_i) \cdot LQ(p_i)$$

Constraints

$$\forall i, j \in [1 \dots K]\ i \neq j, p_i + p_j \text{ if } sim(p_i, p_j) \geq 0.5.$$

$$\sum_i arc_{s,i} = 1 \quad \sum_i arc_{i,e} = 1 \quad \sum_i arc_{i,j} = \sum_i arc_{j,i} \forall j$$

$$\sum_i arc_{i,j} + \sum_i arc_{j,i} = 2p_j \forall j$$

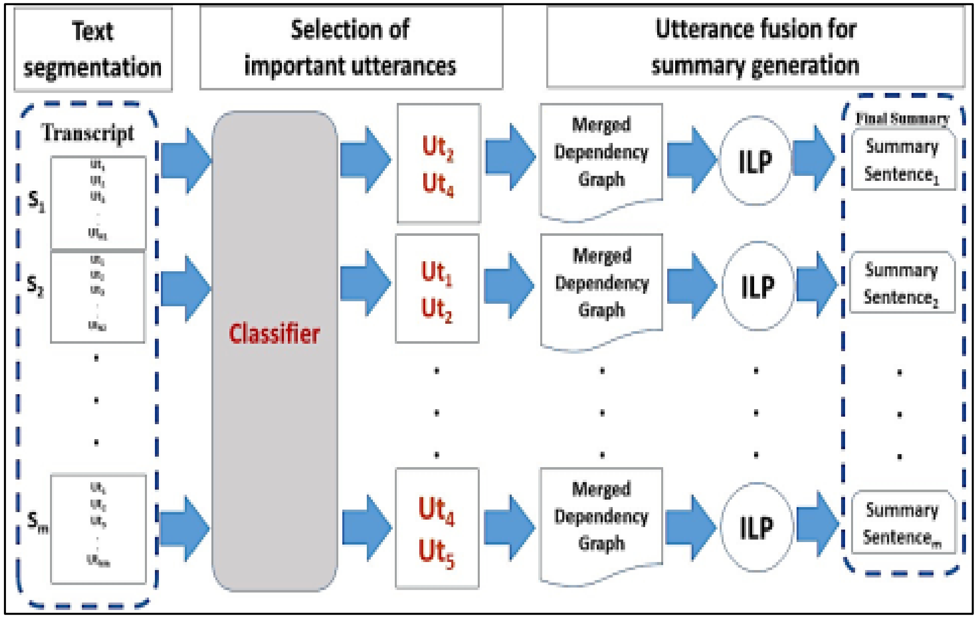
$$\sum_{i=1}^{K} l_i p_i \leq L_{max}$$

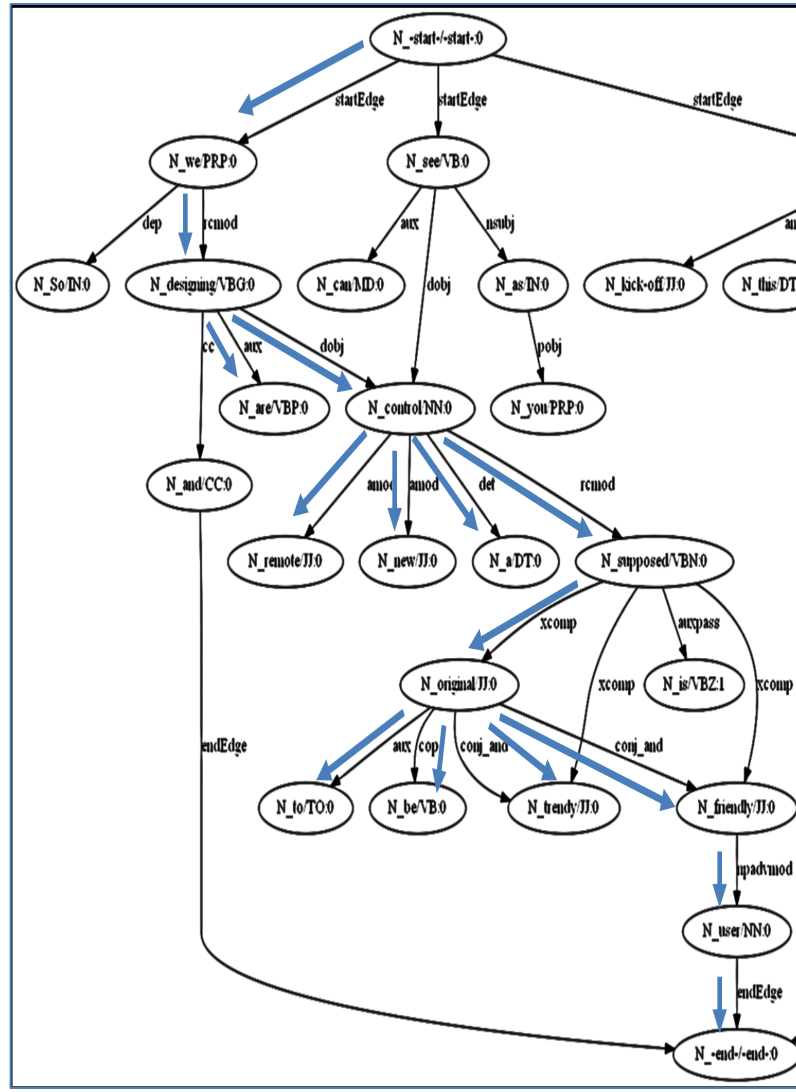
Results: News Summarization

- **ROUGE evaluation** on **DUC 2004** and **2005** datasets – several configurations.
- Our abstractive system performs better than the baselines
- Performance is better than state-of-the-art techniques with several configurations.

DUC-2004			DUC-2005		
	ROUGE-2	ROUGE-SU4		ROUGE-L	ROUGE-SU4
Baselines			**Baselines**		
GreedyKL	0.08658	0.13253	Random	0.26395	0.09066
FreqSum	0.08218	0.12448	Centroid	0.32562	0.11007
Centroid	0.08139	0.12642	LexRank	0.33179	0.12021
TsSum	0.08068	0.12209			
LexRank	0.07796	0.12484			
State-of-the-arts			**State-of-the-arts**		
DPP	0.10079	0.14556	DUC Best	0.34764	0.10012
Submodular	0.09602	0.14227	LSA	0.26476	0.10806
RegSum	*0.09712*	0.13812	NMF	0.28716	0.11278
CLASSY04	0.09168	0.13250	KM	0.29107	0.10806
OCCAMSV	0.09420	0.13105	FGB	0.35018	0.12006
ICSISumm	0.09585	0.13314	RDMS	0.35376	0.12297
CLASSY11	0.08912	0.12779			
Abstractive Systems					
$MD^{Imp}_{LexRank}$ + APO + MSC [Filippova, 2010]	0.09612	0.13911	$MD^{Imp}_{LexRank}$ + APO + MSC	0.35589	0.12211
$MD^{Imp}_{LexRank}$ + MO + ILP	0.09799	0.13884	$MD^{Imp}_{LexRank}$ + MO + ILP	0.35281	0.12107
$MD^{Imp}_{LexRank}$ + APO + ILP	0.10317	0.14218	$MD^{Imp}_{LexRank}$ + APO + ILP	0.35342	0.12117
MD^{Imp}_{CosSim} + MO + ILP	0.09799	0.13884	MD^{Imp}_{CosSim} + MO + ILP	0.35661	0.12331
MD^{Imp}_{CosSim} + APO + ILP	0.10577	0.14215	MD^{Imp}_{CosSim} + APO + ILP	0.35577	0.12298
$MD^{Imp}_{DocsetSim}$ + MO + ILP [**ILPSumm**]	$\mathbf{0.11992}^{\dagger}$	$\mathbf{0.14765}^{\dagger}$	$MD^{Imp}_{DocsetSim}$ + MO + ILP [**ILPSumm**]	$\mathbf{0.35772}^{\dagger}$	$\mathbf{0.12411}^{\dagger}$
$MD^{Imp}_{DocsetSim}$ + APO + ILP	0.11712	0.13567	$MD^{Imp}_{DocsetSim}$ + APO + ILP	0.35679	0.12393

Meeting Summarization (Banerjee et. al, 2015)





Extracted Utterances

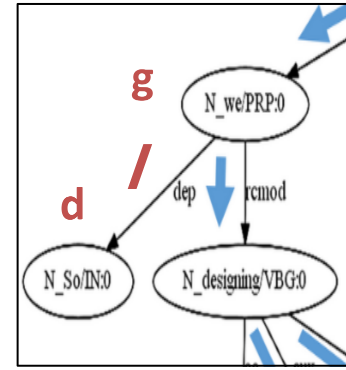
1. *"Um well this is the kick-off meeting for our project."*
2. *"so we're designing a new remote control and um."*
3. *"Um, as you can see it's supposed to be original, trendy and user friendly."*

Anaphora Resolution
Replace "it" with "a new remote control"

We are designing a new remote control supposed to be original trendy and friendly

Mathematical Formulation – Meeting Summarization

- **Objective function:** $\sum_x x_{g,d,l} \cdot p(l \mid g) \cdot I(d) \cdot \frac{p_x}{N}$



Each edge
variable in the ILP

Probabilities
Reuter's corpora (2002)

aux-pass	nsubj-pass	aux	prep_with	agent	prep_in	adv-mod
0.286	0.214	0.214	0.071	0.071	0.071	0.071

Informativeness
Hori and Furui, ICASSP' 2002

$$I(d) = f_s \cdot \log \frac{F_A}{F_d}$$

Position importance
Position of an utterance in the segment

Constraints

$$\forall l \in startEdge, \sum_l x_{g,d,l} = 1,$$
$$\forall l \in endEdge, \sum_l x_{g,d,l} = 1$$
$$\sum_x x_{g,d,l} \leq \gamma$$

$$\sum_{g,d} (x_{g,d,l} + x_{d,g,l}) \leq 1$$
$$\forall l_{out} \in \{aux, cop, det\}, \sum_{u,l_{in}} x_{g,u,l_{in}} - x_{u,d,l_{out}} = 0$$
$$\forall g, l_{out} \in aux \vee cop \vee det, \sum_{l_{out}} x_{g,d,l_{out}} \leq 1$$

Results: Meeting Summarization

❑ **ROUGE Evaluation**

- 300 words limit : Average length of human-written summaries

Method	R-2	R-SU4
Our abstractive model	**0.048**	**0.087**
Our abstractive model (no anaphora resolution)	0.036	0.071
MSC model	0.041	0.079
Extractive model (baseline)	0.026	0.044

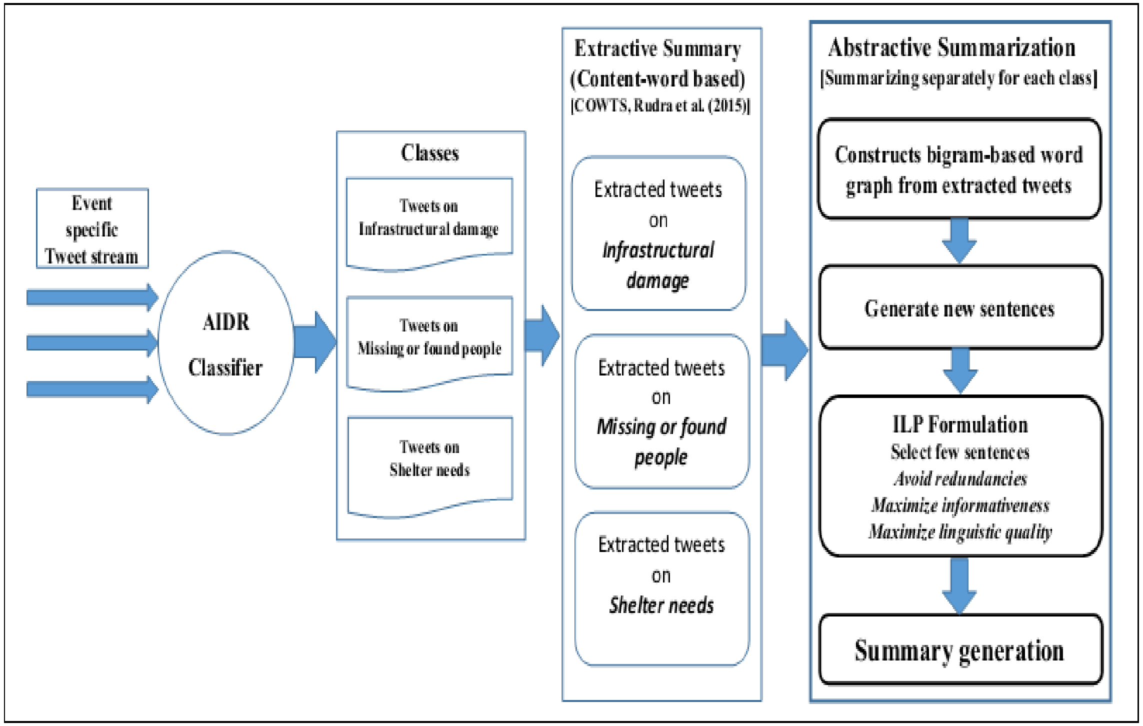
❑ **Readabilit**

Method	Readability score	Log likelihood
Our abstractive model	0.74	**-125.73**
MSC model	0.62	-141.31
Extractive model	0.67	-136.22

MSC Model : Multi-sentence compression [Filippova, COLING' 2010]

This model was adapted in two papers on meeting summarization recently

Twitter Summarization (Rudra et. al, 2016)



Content- word based Summary Quality Optimization

$$max(\sum_{i=1}^{n} LQ(i).I(i).x_i + \sum_{j=1}^{m} y_j) \quad (3)$$

subject to the constraints

$$\sum_{i=1}^{n} x_i \cdot Length(i) \leq L \quad (4)$$

$$\sum_{i \in T_j} x_i \geq y_j, j = [1 \cdots m] \quad (5)$$

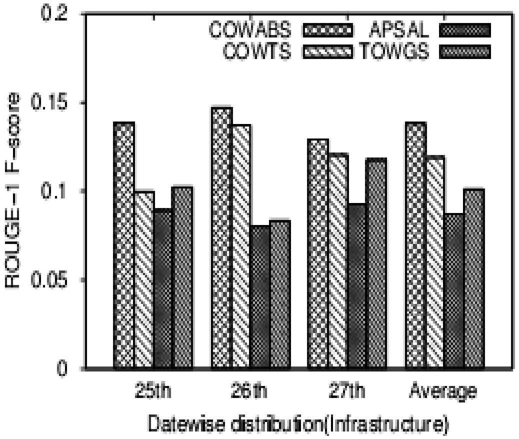
$$\sum_{j \in C_i} y_j \geq |C_i| \times x_i, i = [1 \cdots n] \quad (6)$$

Content words: Numerals, nouns, locations, main verbs

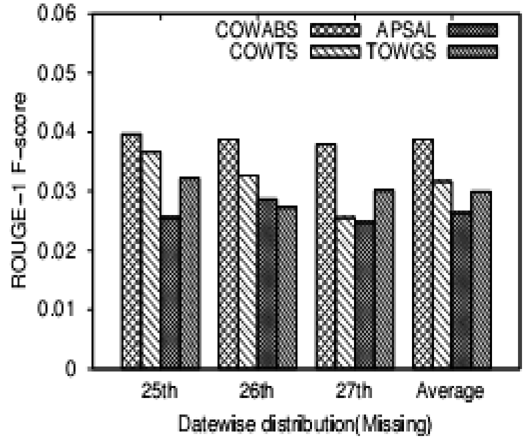
- 5: Content word -> At least One Sentence
- 6: Sentence selected determines content words to be selected

Results: Twitter Summarization

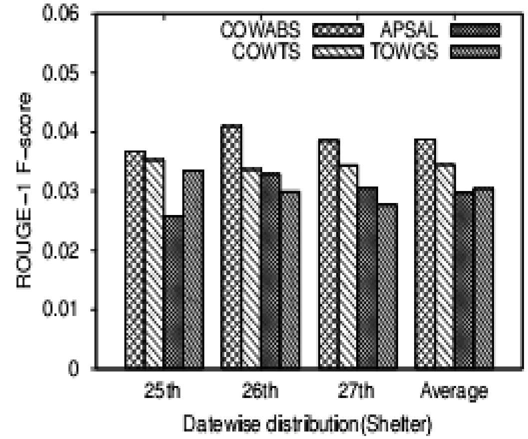
ROUGE-based evaluation



(a) Infrastructure

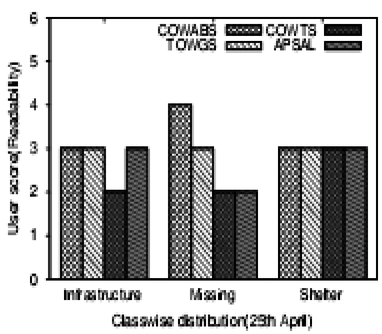


(b) Missing

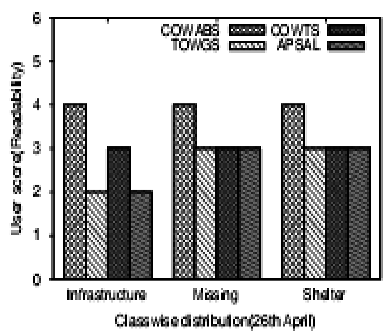


(c) Shelter

Readability evaluation



(a) 25th (readability)



(b) 26th (readability)

An Interdisciplinary Team

Penn State University

- **John Yen**
  University Professor of IST, Penn State
- **Prasenjit Mitra**
  Associate Professor of IST, Penn State

- **Kang Zhao**
  Assistant Professor, Univ. of Iowa
- **Cornelia Caragea**
  Assistant Professor, Univ. of North Texas

American Cancer Society

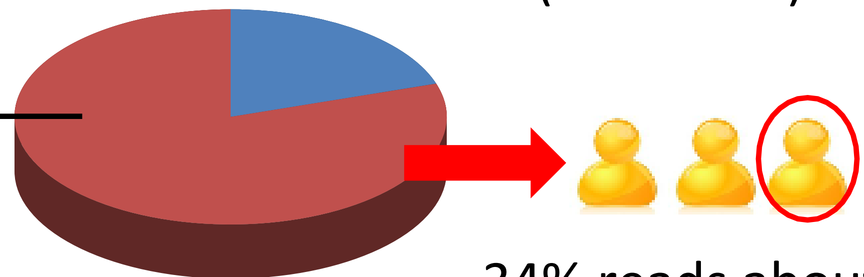
- **Kenneth Portier**
  Managing Director, Statistics & Evaluation Center, Intramural Research

- **Greta Greer**
  Director, Survivor Programs

Online Health Community (OHC)

Adult Internet users in the U.S. (Pew 2012)

80% use Internet for health-related purposes



34% reads about health-related experiences or comments from others

- Cancer Survivors
  - ACS's Cancer Survivors Network
  - CHESS
- Smoking Cessation Community
  - QuitNet
- American Heart Association
  - My Start! Online Community (now Walking Club)
- PatientsLikeMe

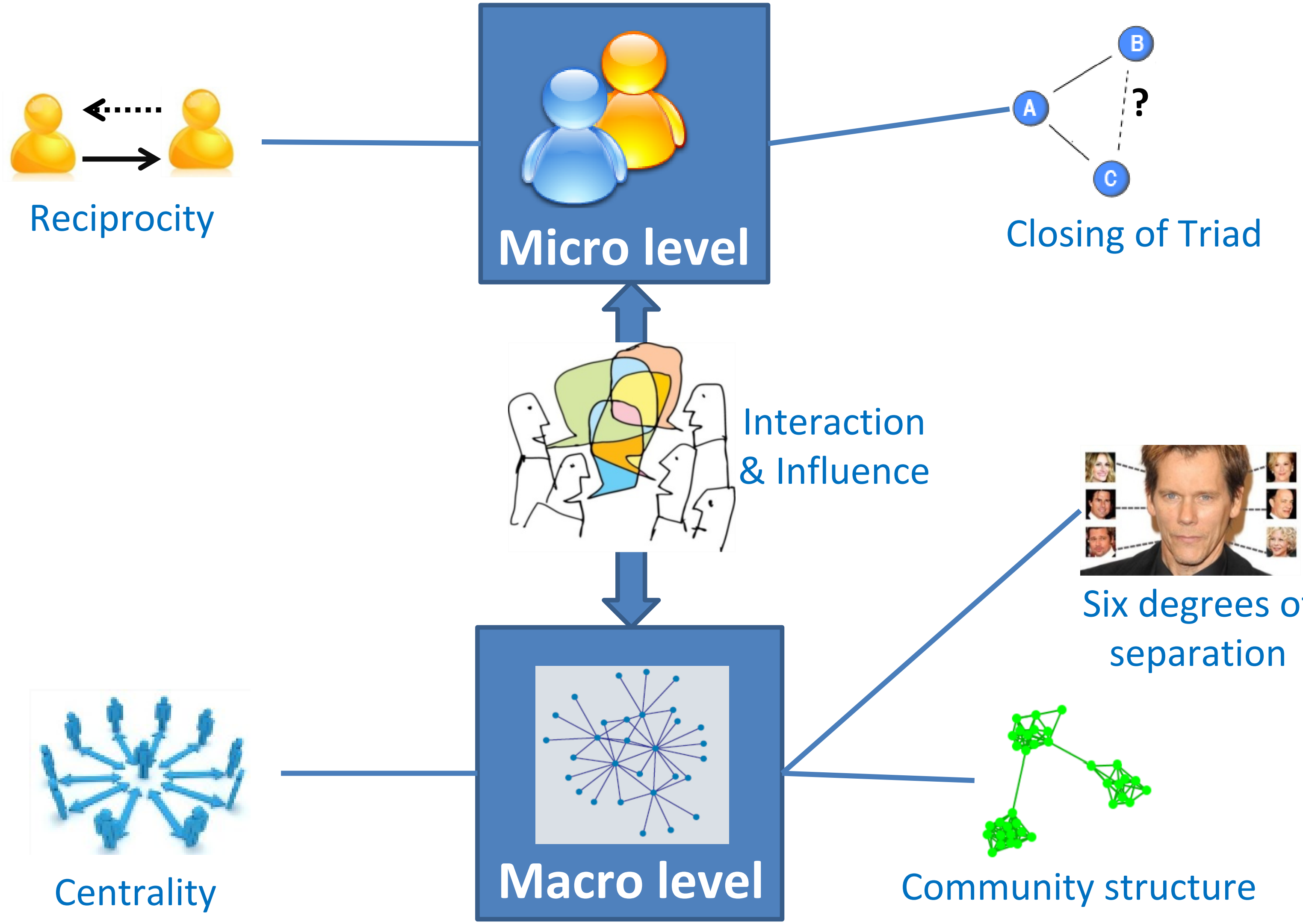
ACS Cancer Survivors Network

- Online communities are an important source of social support for cancer survivors and caregivers.
- The ACS Cancer Survivors Network (CSN), established in 2000, is the oldest and largest of these, with 160, 000+ members.

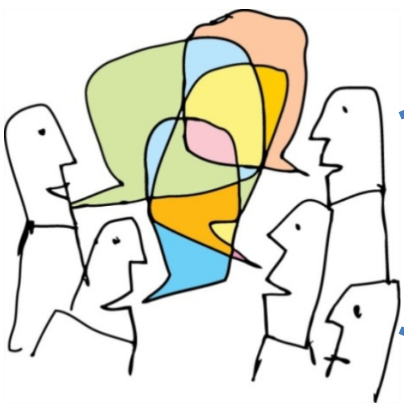
Research Questions

- What are the impacts of participation in ACS Cancer Survivors Network on cancer survivors?
- How are these impacts associated with the topics of discussion?
- How to identify influential users in the online community?

Two Levels of Analysis in Social Media



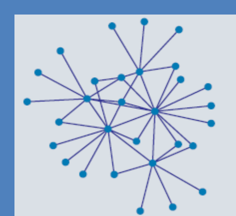
Influence connects the two level



Influence

Micro level

Affects individuals' opinions, emotions, decisions, and actions



Macro level

Social contagion
Influential users

Previous Research on Benefits of OHC

- Increasing social support (Dunkel-Schetter 1984, Rodgers and Chen, 2005)
- Reducing the level of stress, depression, and psychological trauma (Beaudoin and Tao 2008, Winzelberg et al, 2003)
- Helping participants to become more optimistic about disease (Rodgers and Chen, 2005)

Data: Cancer Survivors Network

- Analyzed 48,779 threaded discussions (468,000 posts by 27,173 members) from 2000 to 2010 in CSN’s Discussion Forum
  - Consists of 38 discussion boards
  - 25 are cancer-specific, with the breast cancer and colorectal cancer boards being largest
  - Non-cancer-specific boards cover topics such as humor, caregiving, emotional support and spirituality

CSN Forum Data

- De-identified forum posts
  - Peer to peer interaction (no physicians/doctors)

TABLE I
SUMMARY OF STATISTICS IN THE FORUM DATASET

	Mean	Median	Maximum
Number of posts by a user	17.25	2	5,607
Number of replies per thread	8.7	6	442
Life span of a thread	1,725 hrs	58 hrs	87,846 hrs

Power-law distributions

- The number of posts per user
- The number of posts per thread
- The life spans of threads

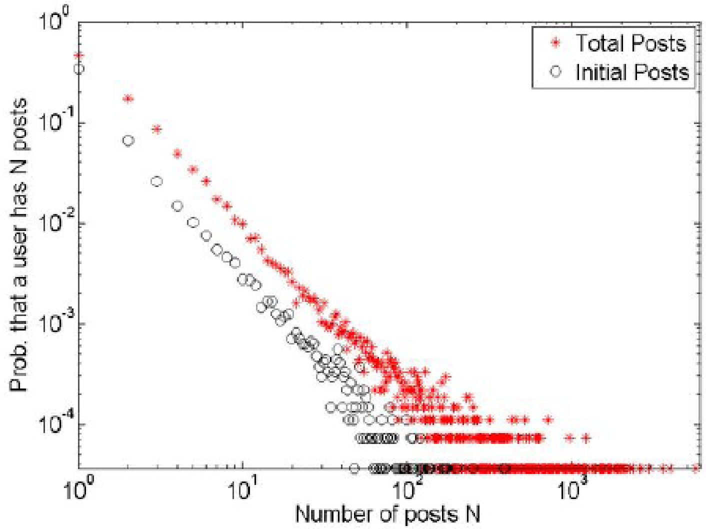


Fig. 1. Distributions of users' contribution (initial posts and total posts) to the online forum

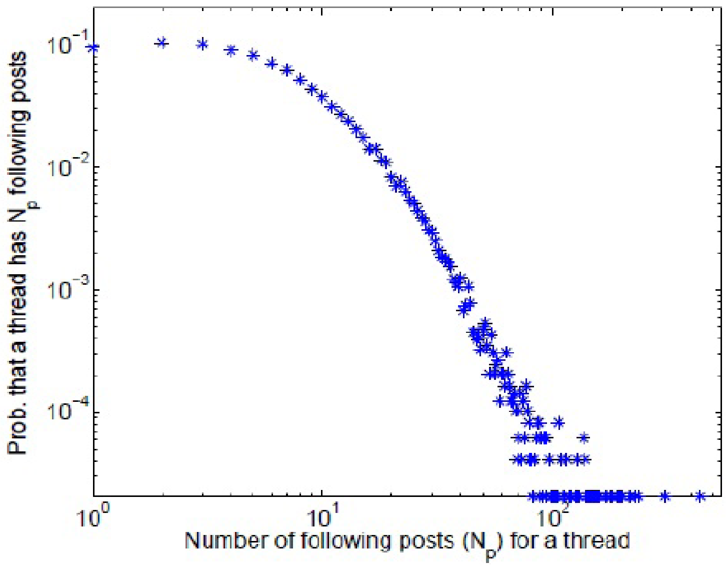


Fig. 2. Distribution of the number of replies in threads

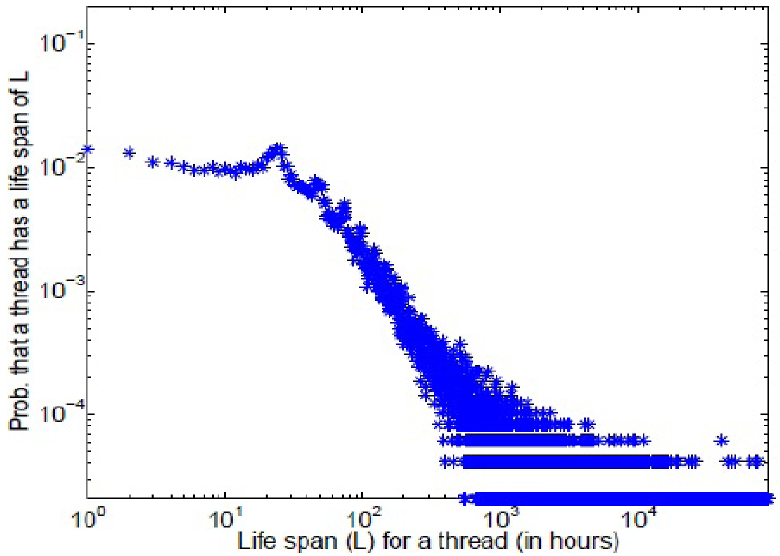
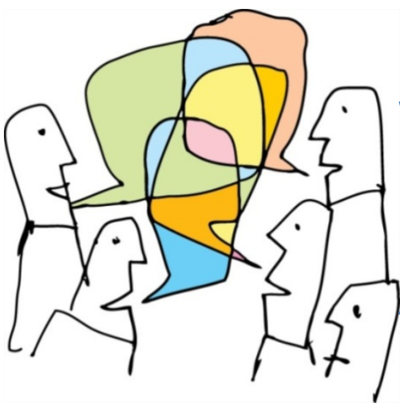


Fig. 3. Distribution of the time span of threads

A Framework for Analyzing Influence in Social Media



Influence

3. Identify influential posts for the changes

Micro level

1. Assess individual's state

2. Identify state changes in a conversation

Macro level

4. Identify influential users in a community

An example of psychosocial support

Initial post of A

The initial post

User A: My mom was diagnosed with Lung Cancer that the doctor described as "advanced" ... She was emotionally strained .. this is very upsetting to me and my family ...

One reply post

A responding reply

User B: ..., I had been diagnosed with lung cancer... The tumor has shrunk ... and is dormant hopefully forever.. keep good spirits and give her a lot of support ... I feel that she will beat this. ... I'd be glad to help.

Self-reply of A

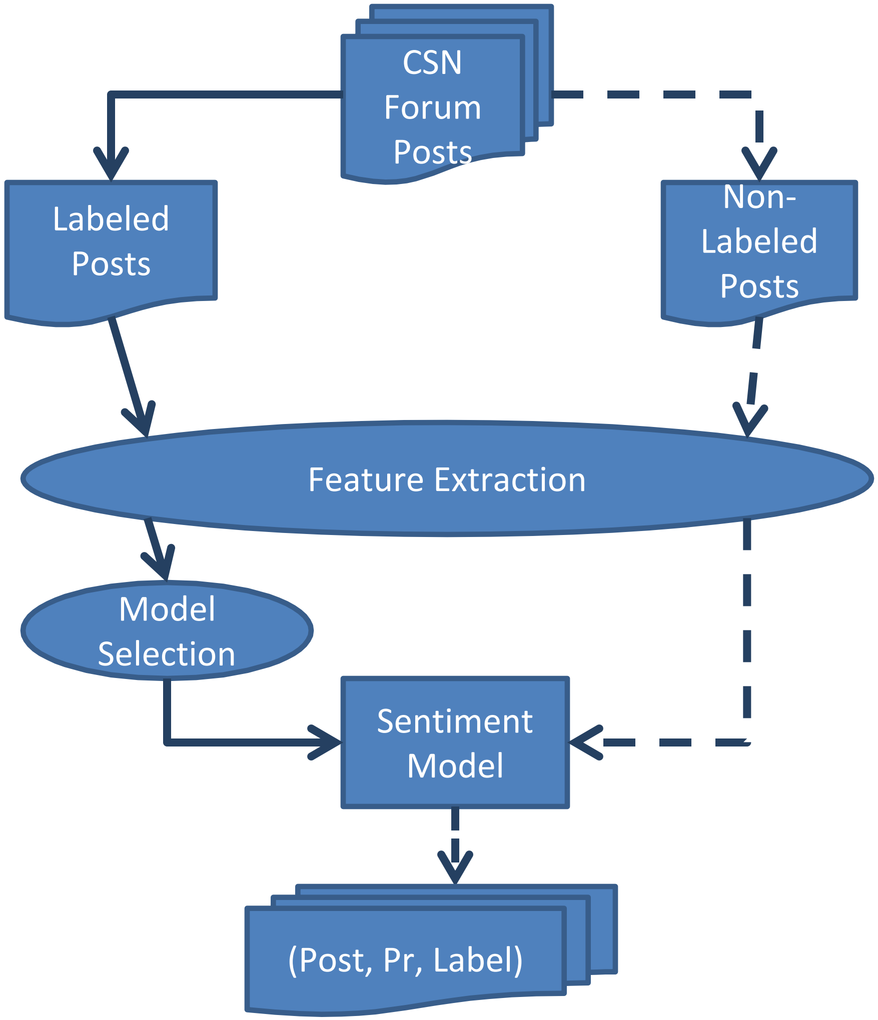
A self-reply

User A: .. thank you so much for your response. That was encouraging.. I showed her your message .. gave her hope... a sign of relief on her face...

Sentiment Analysis

- Bag-of-Word sentiment analysis
  - Associate individual words in a post with positive/negative sentiment
  - Associate individual words with strength of positive/negative sentiment
- Part-of-Speech Tagging
  - Use sentence structure to extract words related to sentiment as well as <u>what the sentiment is about</u> (Liu et al., 2005)
  - Useful for sentence-level analysis
- Unsupervised Learning (Liu, 2010)
- Our Approach: Supervised Learning

Automatic Sentiment Classifier



- Model training
  - Manually labeled 298 posts (Positive or Negative)
  - Use the labeled posts to train a sentiment classifier
  - Feature extraction
  - If the output of the sentiment classifier is greater than 0.5, the post is Positive; otherwise, the post is Negative
  - Evaluate the model using un-trained labeled posts (cross validation)
- Model prediction
  - Apply the model to label the sentiment of unlabeled posts of the entire community

Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E Greer, Kenneth Portier, *"Get Online Support, Feel Better-Sentiment Analysis and Dynamics in an Online Cancer Survivor Community."* in: *Proceedings of* the Third IEEE International Conference on Social Computing, Boston, Massachusetts, USA, 2011.

Features from modified sentiment word-lists

Features

Novel features introduced through insights

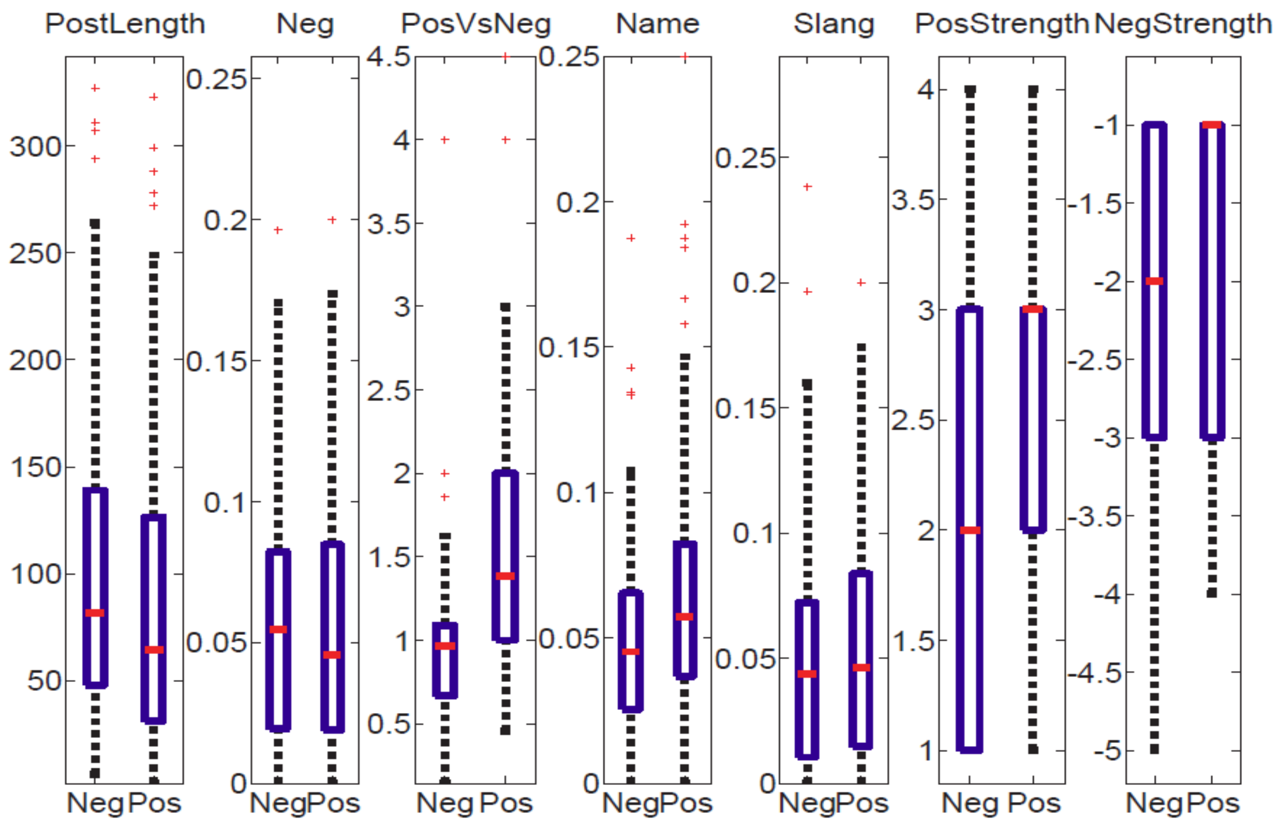
Feature	Definition
PostLength	The number of words
Pos	NumOfPos/PostLength, where NumOfPos is the number of positive words/emoticons
Neg	NumOfNeg/PostLength, where NumOfNeg is the number of negative words/emoticons
Name	NumOfName/PostLength, where NumOfName is the number of name mentioned
Slang	NumOfSlang/PostLength, where NumOfSlang is the number of Internet slang
PosStrength	Positive sentiment strength [22]
NegStrength	Negative sentiment strength [22]
PosVsNeg	(NumOfPos+1)/(NumOfNeg+1)
PosVsNegStrength	PosStrength/NegStrength
Sentence	The number of sentences
AvgWordLen	The average length of words
QuestionMarks	The number of question marks
Exclamation	The number of exclamations

Sentiment Model

Model	ROC Area	Accuracy
AdaBoost	0.832	79.2%
Logistic Regression	0.832	77.5%
LogitBoost	0.816	76.8%
BayesNet	0.802	74.2%
Bagging	0.794	73.5%
Neural Networks	0.785	73.8%
Decision Tree	0.782	77.2%
SVM	0.658	75.2%

- 10-fold cross-validation
- Best model: AdaBoost-based sentiment model
  - The false positive rate of AdaBoost is 0.152
  - The false negative rate is 0.33
- Best feature set for AdaBoost
  - PostLength, Neg, PosVsNeg, Name, Slang, PosStrength, NegStrengthg

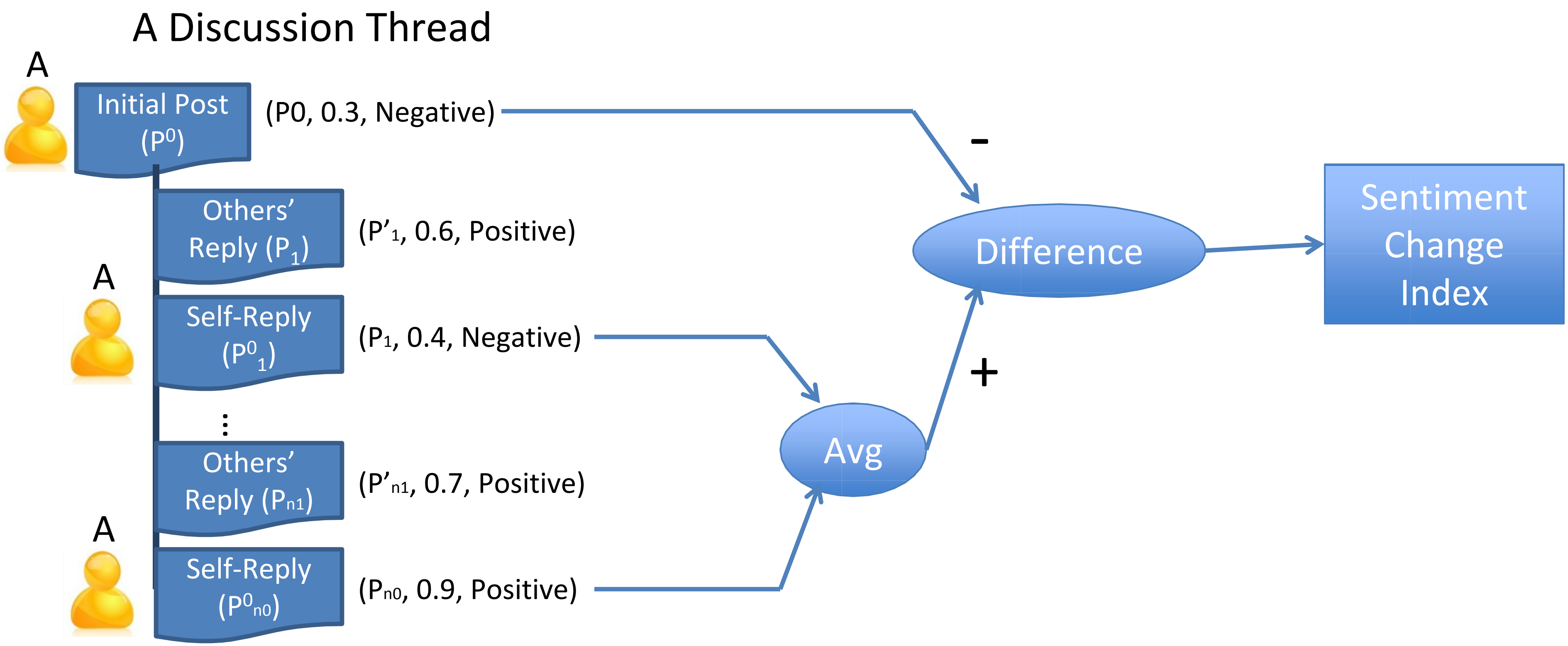
The Best Feature Set



LOO Feature	ROC Area	Decrease from the full model
PosStrength	0.774	0.054
PosVsNeg	0.804	0.024
Neg	0.813	0.015
Slang	0.813	0.015
NegStrength	0.813	0.015
PostLength	0.819	0.009
Name	0.82	0.008

Feature	ROC Area
PosStrength	0.696
PosVsNeg	0.694
Name	0.572
PostLength	0.545
NegStrength	0.544
Slang	0.527
Neg	0.459

Sentiment Change of Thread Initiators

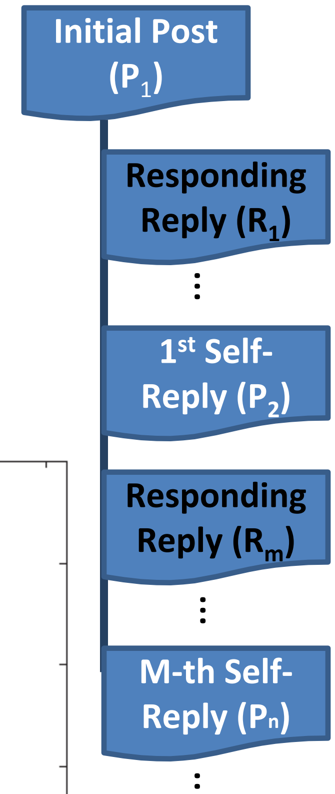


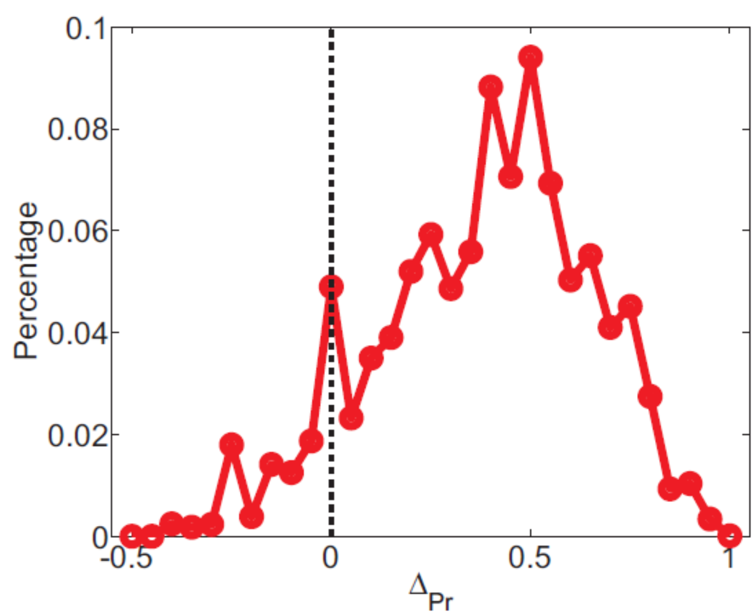
•Sentiment Change Indicator: The difference between the average sentiment of the originators' self replies and the initial sentiment of the thread originator

$$\Delta_{Pr} = \sum_{i=1}^{n_0} Pr(P_i^0)/n_0 - Pr(P^0),$$

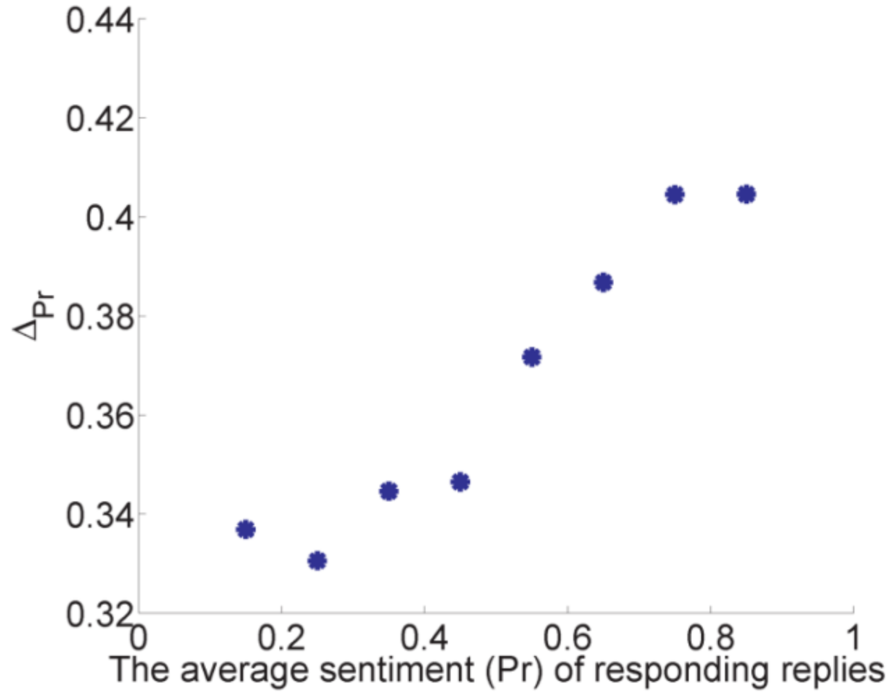
Distribution of The Sentiment Changes

- 75% of the originators with negative initial posts became more positive in self-replies





What correlates with the sentiment changes?

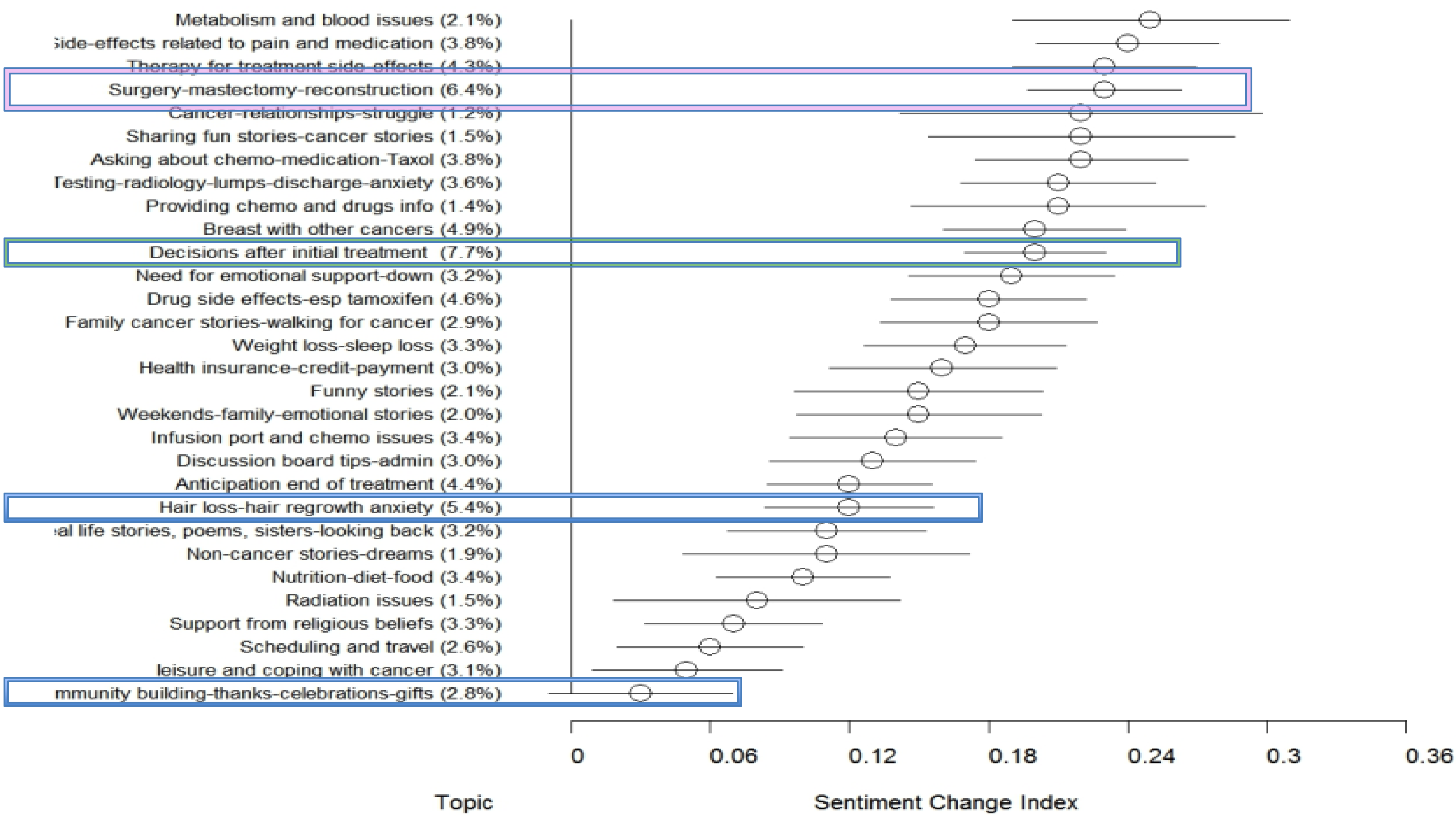


The more **positive** are the sentiment of replies from others, the more **positive** is the sentiment change of the thread originator.

Do topics of initial post sentiment changes?

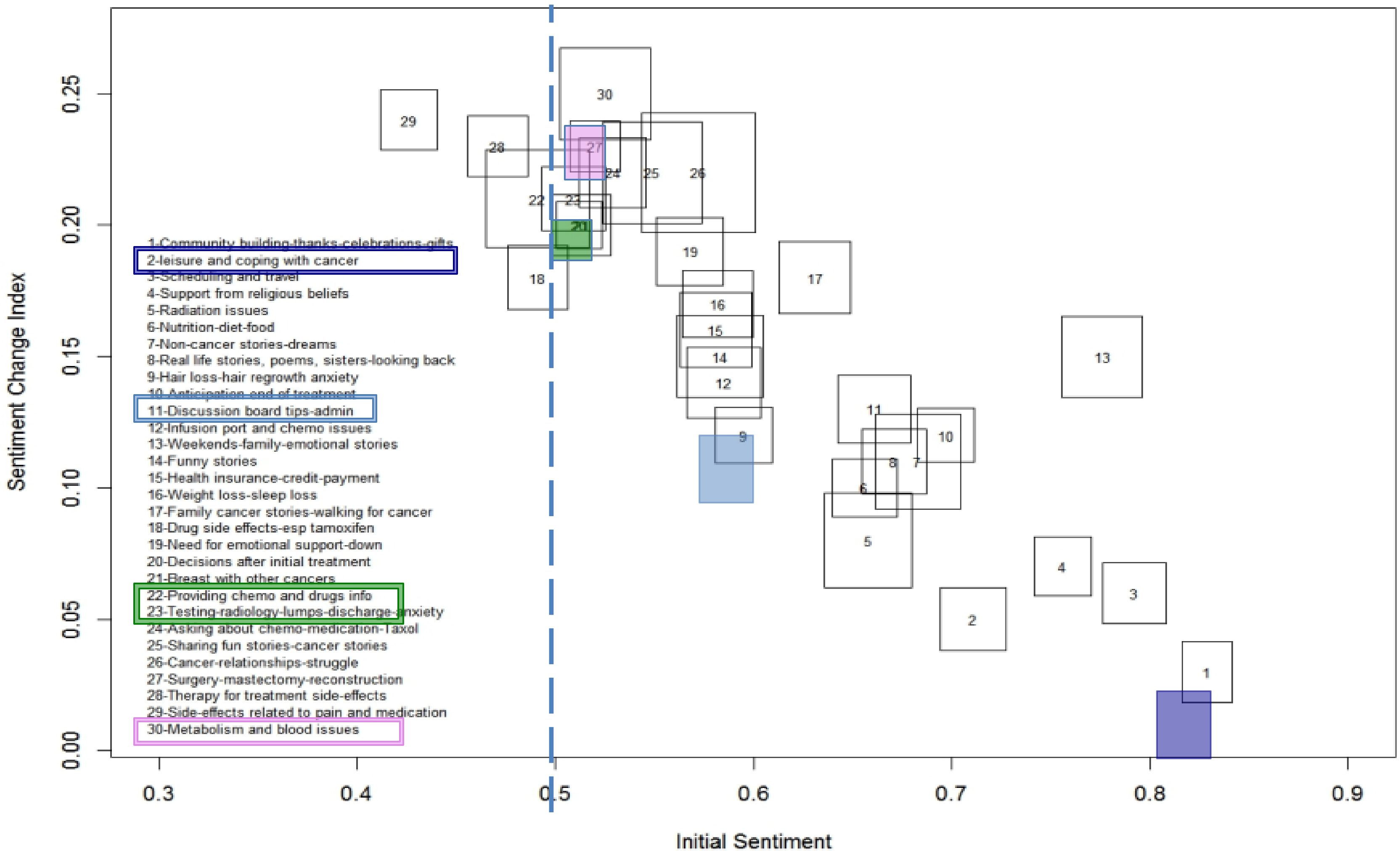
- Cluster initial posts of Breast Cancer subforum and Colon Cancer subforum (2005-2010) into 30 topics using LDA (unsupervised learning)
- Assign each initial post to the topic with highest probability.
- Calculate the average and variance of sentiment changes for all initial posts in each topic.

Sentiment changes are higher for breast cancer topics related to treatments and side effects



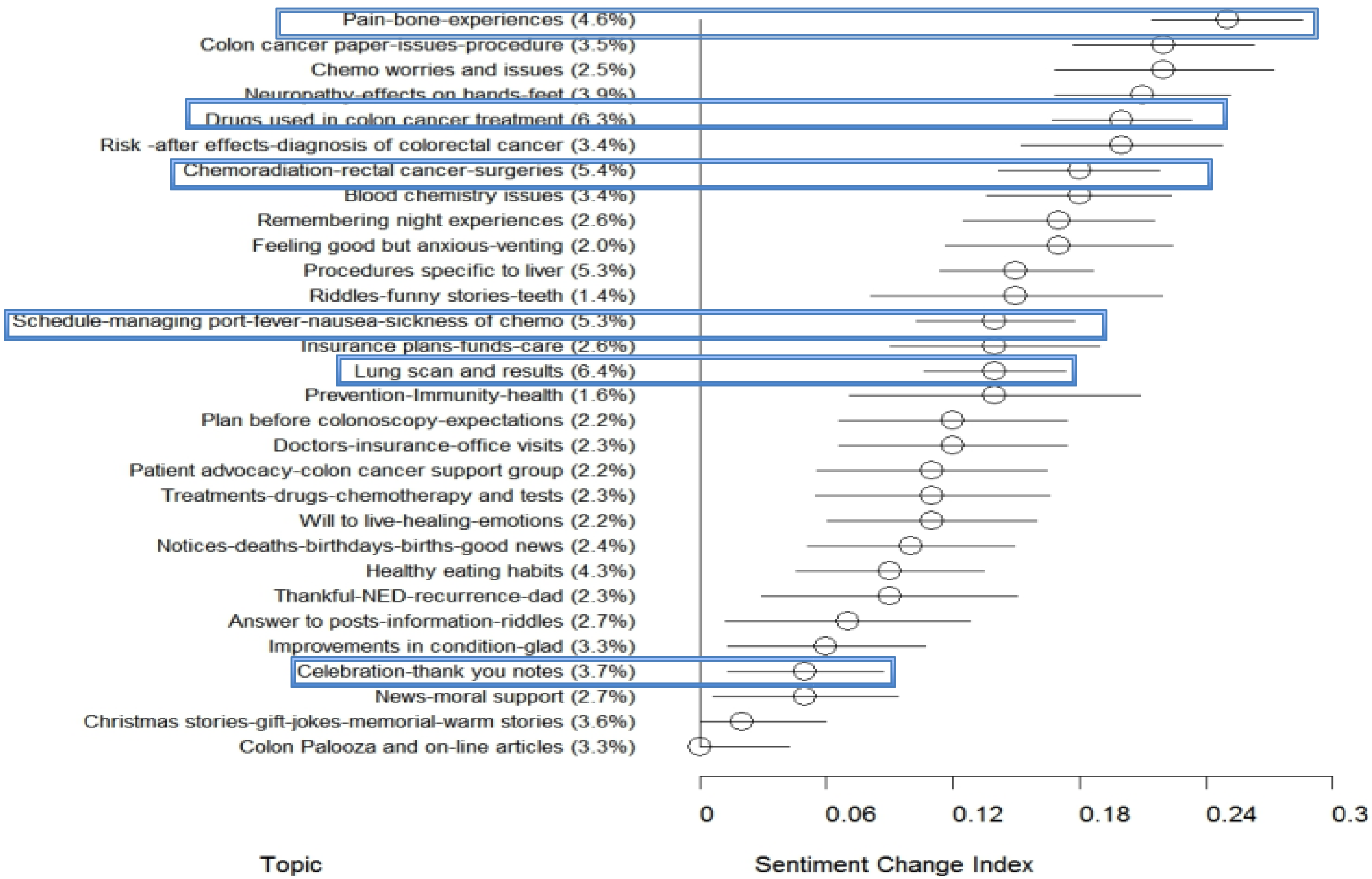
Portier, K., Greer, G., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P., and Yen, J., "Understanding topics, sentiment, and influence in an online cancer survivor community", JNCI (pending).

Sentiment changes are higher for breast cancer topics with lower initial sentiment.



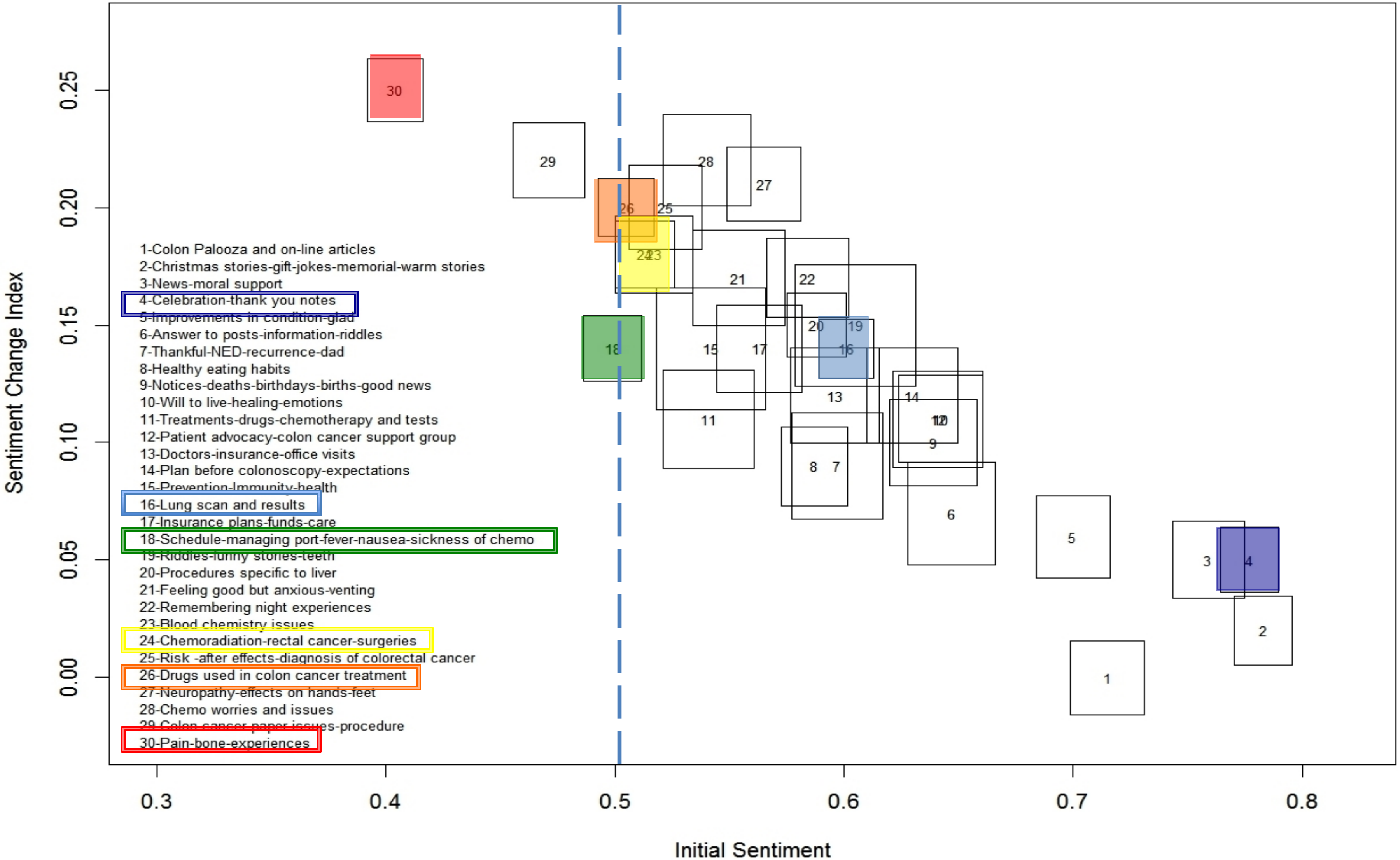
Portier, K., Greer, G., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P., and Yen, J., "Understanding topics, sentiment, and influence in an online cancer survivor community", JNCI (pending).

Sentiment changes are higher for colon cancer topics related to treatments and side effects



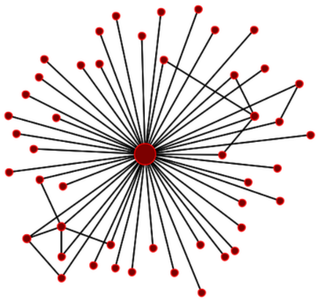
Portier, K., Greer, G., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P., and Yen, J., "Understanding topics, sentiment, and influence in an online cancer survivor community", JNCI (pending).

Sentiment changes are higher for colon cancer topics with lower initial sentiment.

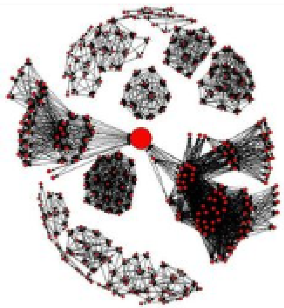


Approaches to Find Influential Users

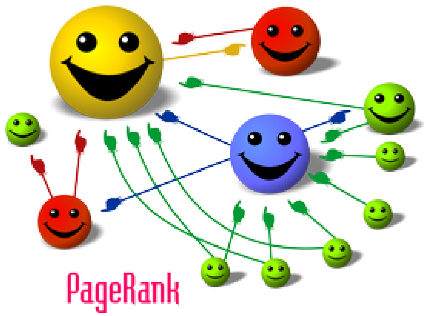
Degree-based centrality



Path-based centrality

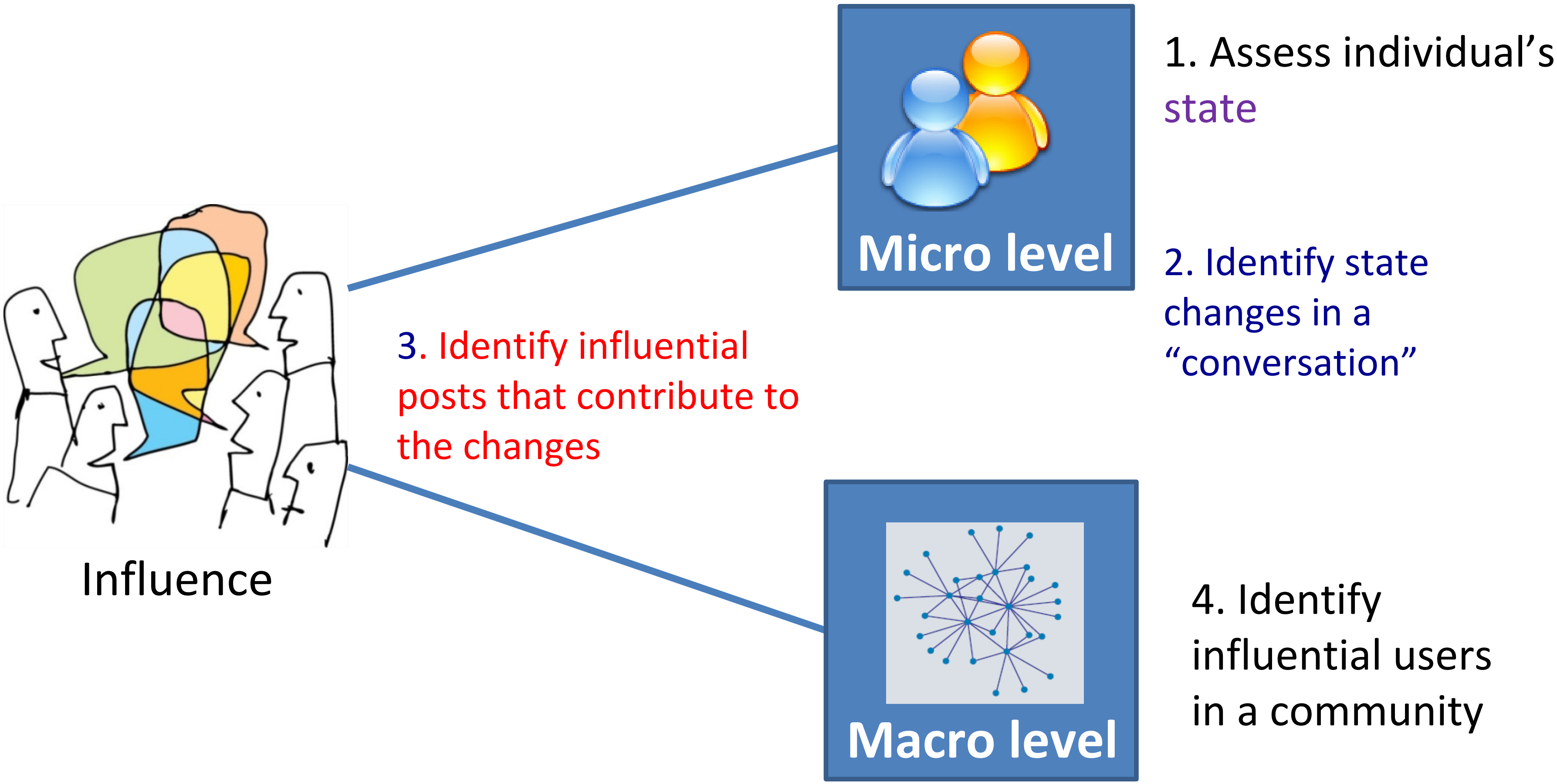


Link-based centrality

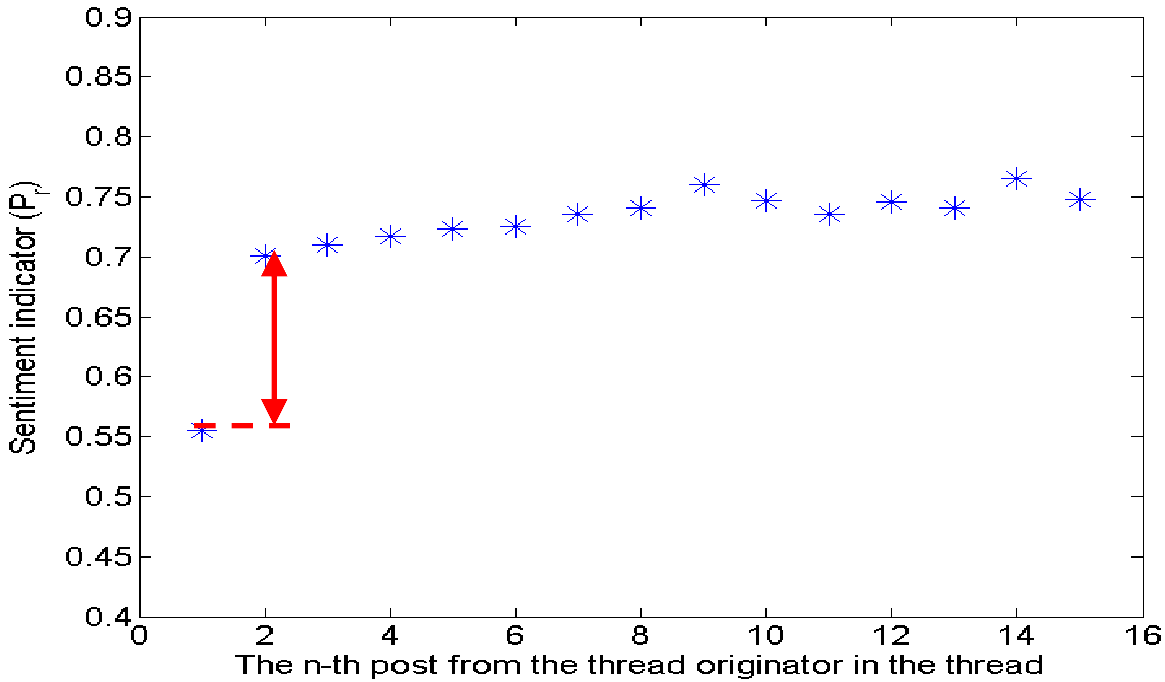


- Model-based influence maximization
  - Needs a <u>computational influence model</u>
    - Threshold-based Model
    - Independent Linear Cascade Model

A Framework for Analyzing Influence in Social Media

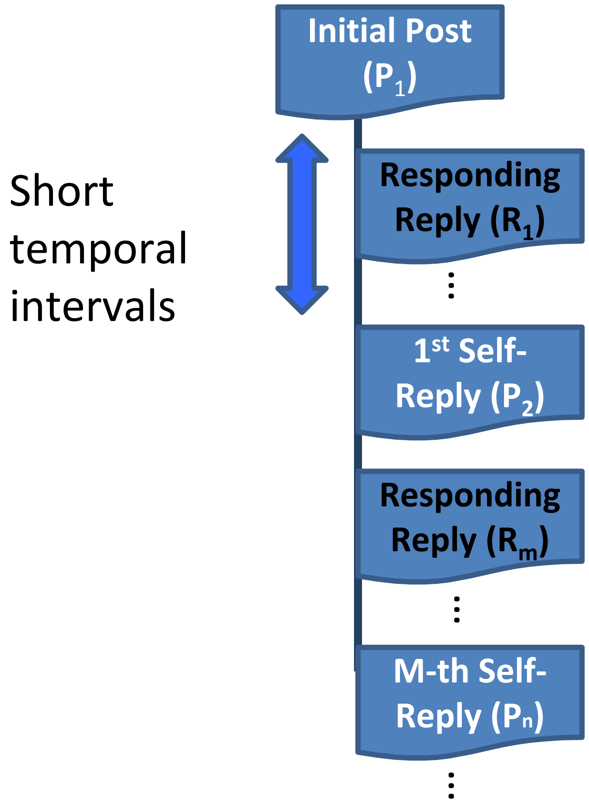


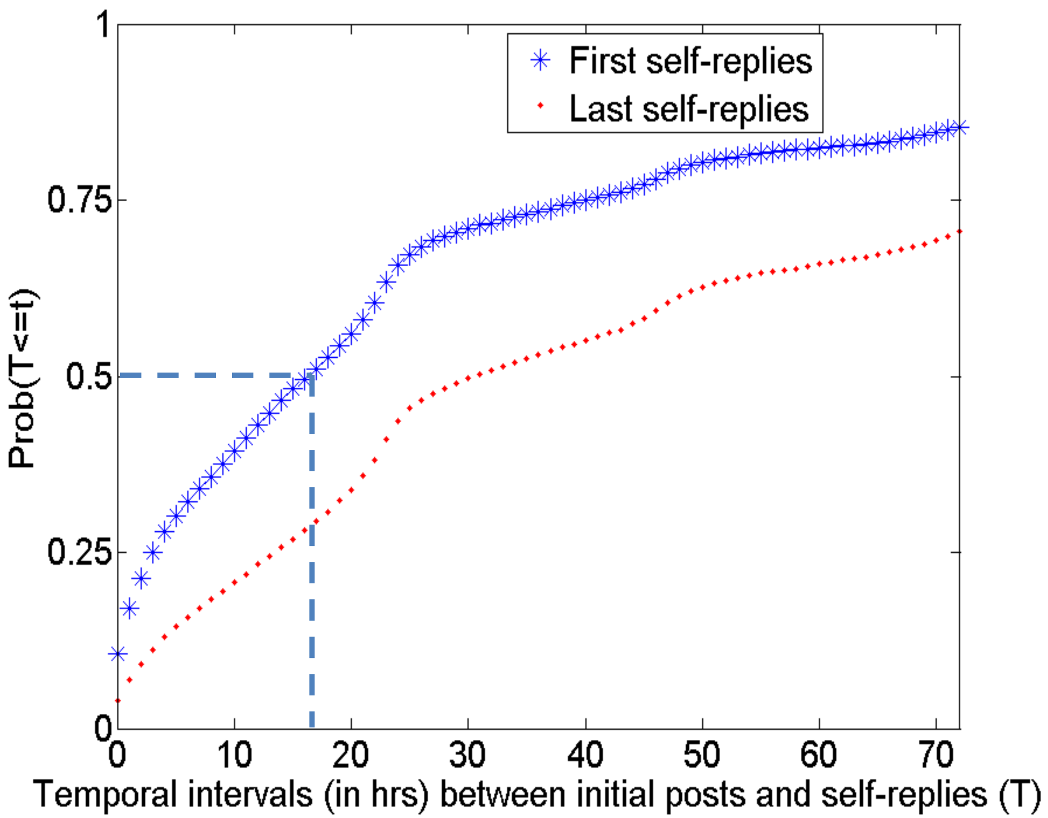
How to identify posts that influence the sentiment change of the thread initiator?



- The sentiment change mostly occurs at the first self-reply.
- After the first self-reply, the sentiment of the thread originator, on the average, does not change much.

This change occurs within a short period





- Most of the sentiment change occurs between the original post and the first self-reply.

Influential posts

- They were posted before the first self reply
- Their sentiments are aligned with the change of the thread initiators
    - If the change is positive, we look for posts with positive sentiment
    - If the change is negative, we look for posts with negative sentiment

Sentiment-based Influential User Ranking

- The total number of influential posts (IRR)
- The ranking by IRR has higher top-K recalls than rankings by many other metrics.

Compare Top-K recalls			
Metrics	**K=50**	**K=100**	**K=150**
Total number of threads initiated	0.342	0.439	0.585
Total number of posts	0.415	0.707	0.781
In-degree in the social network	0.317	0.512	0.610
Out-degree in the social network	0.390	0.659	0.780
Betweenness in the social network	0.293	0.366	0.488
PageRank in the social network	0.390	0.561	0.732
Total number of IRR	0.511	0.732	0.805

Compare Two Approaches for Identifying Influential Users

- Approach 1: Sentiment-based Approach
- Approach 2: Ensemble Classifier

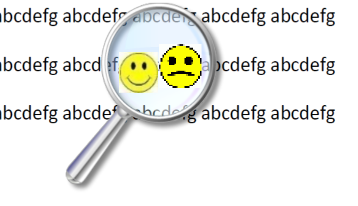
Kang Zhao, Baojun Qiu, Cornelia Caragea, Dinghao Wu, Prasenjit Mitra, John Yen, Greta E Greer, Kenneth Portier. *"Identifying Leaders in an Online Cancer Survivor Community."* In: *Proceedings of the 21st Annual Workshop on Information Technologies and Systems (WIST 2011)* , Shanghai, China, December 3-4, 2011.

Features of Classifier (60+)

- Contribution features
  - The numbers of posts/threads
  - The length of posts
  - The time span of one's activities
  - …

- Centrality features
  - A post-reply network among users
  - 27k+ nodes and 163k+ edges
  - In/out-degree, betweenness, PageRank

- Semantic features
  - Appearance of words with positive/negative sentiment in a user's posts
  - Use of slang and emoticons
  - …



Evaluation

- What about those who are identified as influential users but are not in the list of 41 nominated users?
  - We asked domain experts to evaluate a new list of 130 potential influential users identified by the ensemble classifier
  - 85 users were endorsed as influential users
  - A total of 126 influential users identified
- A more challenging task for our metric
  - How does the performance of the sentiment-based ranking compare to the classifier?

Compare the sentiment-based approach with the ensemble classifier

- The IRR ranking (based on a single metric) outperforms the original ensemble classifier that uses 60+ metrics
- Incorporating IRR into the classifier further improves the classifier's performance.

Top-K recalls and precisions	K=50		K=100		K=150	
	Recall (max =0.397)	Prec.	Recall (max =0.794)	Prec.	Recall	Prec. (max= 0.840)
The IRR Ranking	0.349	0.880	0.627	0.790	0.762	0.640
The ensemble classifier without IRR	0.278	0.700	0.532	0.670	0.698	0.587
The new ensemble classifier with IRR	0.373	**0.940**	0.579	0.730	**1.000**	0.840

Summary

- Sentiment analysis of online interactions enable us to extract meaningful sentiment information, and their dynamics, from online interactions in social media.
- These meta-level information contributes to
  - better understanding about social influence, topics, community leadership;
  - design interventions and policy to better address needs of cancer survivors.

Current Research

Summarizing Situational and Topical Information During Crises

Koustav Rudra,Niloy Ganguly, Pawan Goyal,
Indian Institute of Technology, Kharagpur, WB, India
Muhammad Imran
Qatar Computing Research Institute, Doha, Qatar
Siddhartha Banerjee, and Prasenjit Mitra
The Pennsylvania State University, University Park, PA 16802, USA

Use of Twitter during Calamities


The Telegraph

Home News World Sp

Technology News Techn

HOME » TECHNOLOGY » TWIT

Japan earthqual

Twitter, Facebook and o
millions of people caught

Twitter homepage Photo: NetPhoto

By Harry Wall
4:15PM GMT 13 Ma

Follow 21.5K

Firstpost > India

Hyderabad blasts: Twitter abuzz with condolences, anger

by FP Staff Feb 22, 2013

Follow @firstpostin

#Government of India #Hyderabad #Hyderabad blasts #TweetFeed #Twitter

Share 6 0 0 Like 1 Email 1 Comment Print

Twin blasts in Hyderabad on Thursday evening left 14 dead and more than a hundred injured. As terror struck India again, the whole country was left in shock.

Netizens took to twitter to express their shock and condolences over the blasts that rocked the Dilsukh area of Hyderabad. Some expressed anger and even criticised Union Home Minister Sushil Kumar Shinde's visit to the site.

Objective

- **Broad Objective:** Improve utilization of OSNs (Twitter) in the aftermath of natural/man-made calamities, through retrieval of ***useful*** information.

- Need for summaries
  - http://reliefweb.int/report/nepal/nepal-earthquake-2015-post-disaster-needs-assessment-executive-summary
- Need for finer-grained information
  - Which bridges are down? Which roads closed?

Dataset

- Events considered:
  - Nepal Earthquake (25th – 27th April, 2015) classified data using **AIDR**
  - Infrastructure and utilities (16,842)
  - Missing or trapped (10,751)
  - Shelter and supply (19,006)

Concept based Extractive Abstractive Summarization (CONABS)

Summarization

- Extractive

  #PM chairs follow-up meeting to review situation following #earthquake in #Nepal @PMOIndia #nepalquake. @SushmaSwaraj @MEAcontrolroom

- Abstractive
  - India provides aid to Nepal earthquake victims.

Characteristics

- Information Coverage
- Redundancy
- Readability
- Real-time

Two-stage

- First step: extractive
- Second step: abstractive

Summarizing situational updates

- Some particular types of words play an important role in disaster
- Consider specific types of terms (Content words)
  - Numerals (number of casualties, helpline nos.)
  - Nouns (names of places, important context words like people, hospital)
  - Main Verbs (killed, injured, stranded etc.)

Concept & Event extraction

- Nouns represent concepts and verbs represent events
- Micro level information consists of two core nuggets – a noun part, a verb part
- Develop undirected weighted graph among nouns
- Edge weights represent semantic similarity between two nouns
- Cluster similar nouns like ‘airport’ and ‘flight’
- Each cluster represents one **concept**
- Similarly each verb cluster represents one **event**

Objective

- Reducing redundancies in final summary
- Combining information from similar tweets

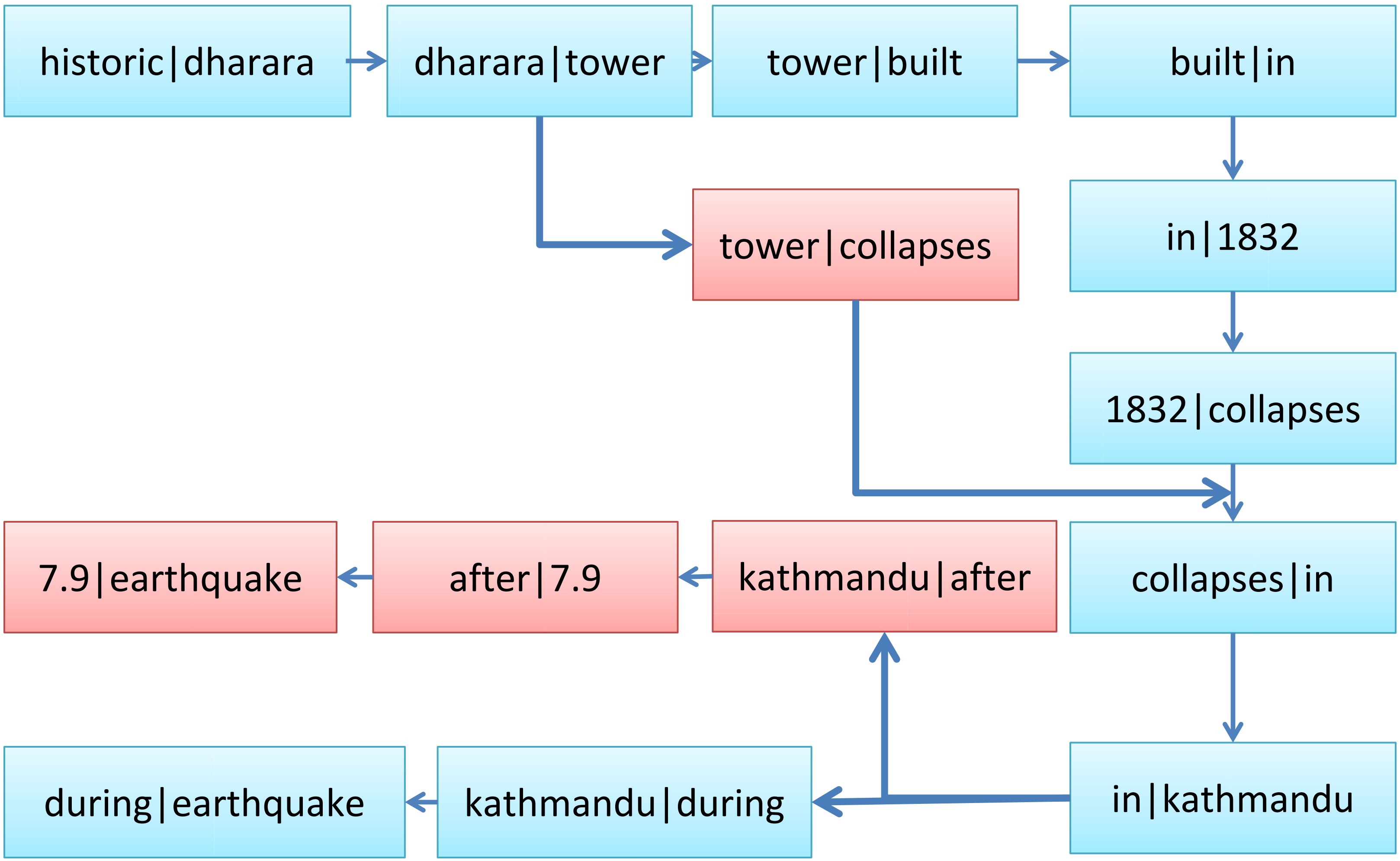
Dharara Tower built in 1832 collapses in Kathmandu during earthquake.
Historic Dharara Tower Collapses in Kathmandu after 7.9 Earthquake.

Dharara tower built in 1832 collapses in Kathmandu after 7.9 earthquake

Approach

- Generate a word graph where nodes are bigrams [deal with informal nature of tweets]
- Generate sentences from the word graph
- **Challenge: Maintaining coherence and readability**
  - Favor sentences generated from a combination of 2-3 tweets
  - Intra-sentence similarity
  - Linguistic quality
  - ILP model combining above factors

- Dharara Tower built in 1832 collapses in Kathmandu during earthquake.
- Historic Dharara Tower Collapses in Kathmandu after 7.9 Earthquake.



ILP based formulation

Parameters

- Score of sentences/generated paths (I(s))
  - Textrank/ Centroid score
- Linguistic quality(LQ(s))
  - Trigram language model
  - $LQ(s) = 1/(1-ll(w_1,w_2,\ldots,w_q))$
  - $ll(w_1,w_2,\ldots,w_q) = 1/L \log_2 \prod^{q}_{t=3} P(w_t|w_{t-2}w_{t-1})$

ILP based solution

$\max(\sum_{i=1 \ldots n} I(i).LQ(i).x_i + \sum_{j=1 \ldots m} y_j)$	x_i, y_j binary variable x_i tweet-path indicator, y_j content word indicator

Constraints

$\sum_{i=1 \ldots n} x_i * \text{Length}(i) \leq L$	Length(i) = number of words in tweet i L = required summary **word length limit**
$\sum_{i \in T_j} x_i \geq y_j \quad j = [1 \ldots m]$	T_j = set of paths where content word j is present If y_j is selected then at least one tweet covering that word is also selected

Now content words are

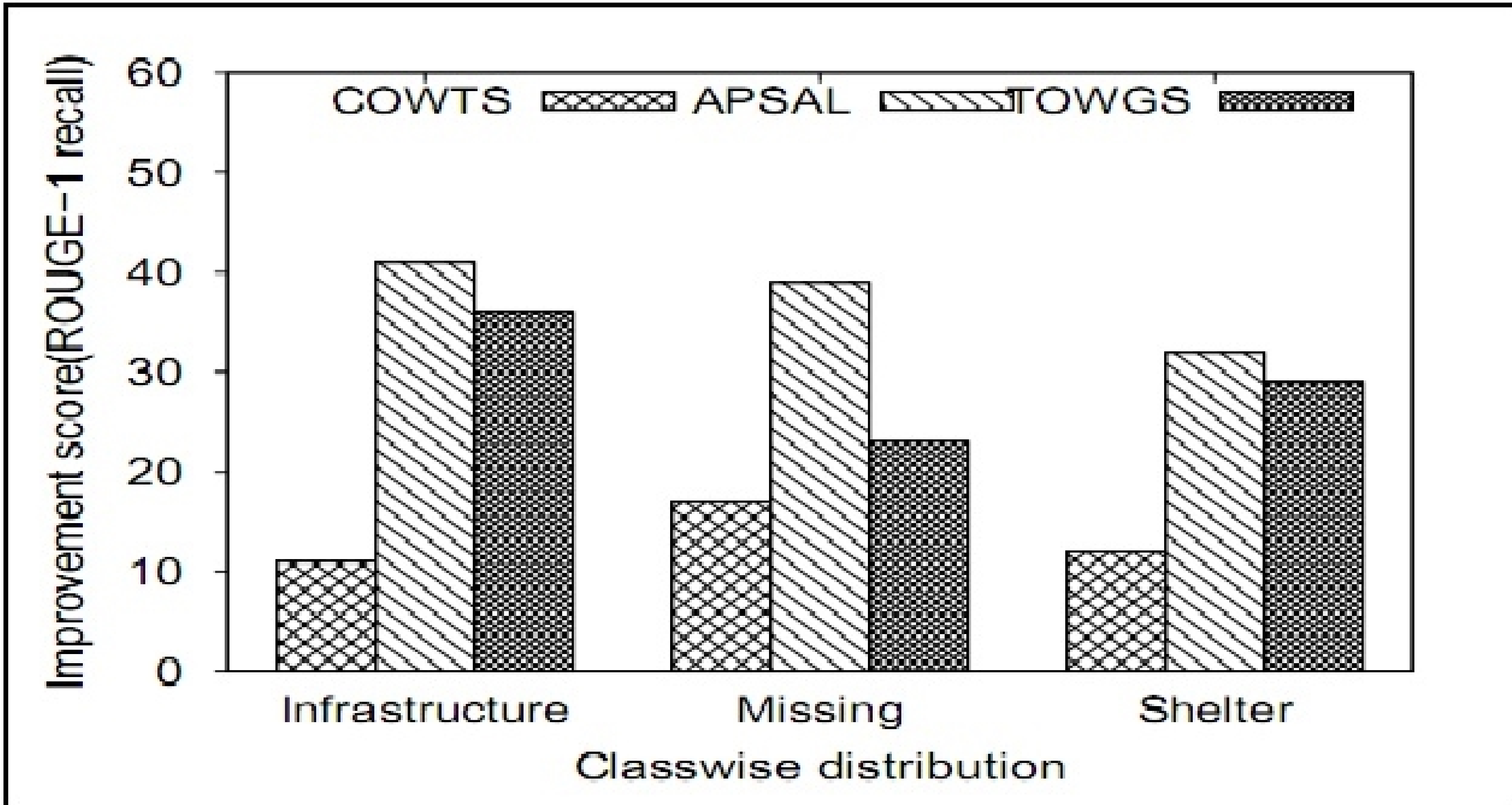
1. Numerals
2. Concepts
3. Events

Baselines

- **COWTS:** runtime content-word based tweet stream summarization algorithm [Rudra 2015]
- **APSAL:** affinity clustering based summarization technique [Kedzie 2015]
- **TOWGS:** runtime bigram based abstractive summarization algorithm [Olariu 2014]

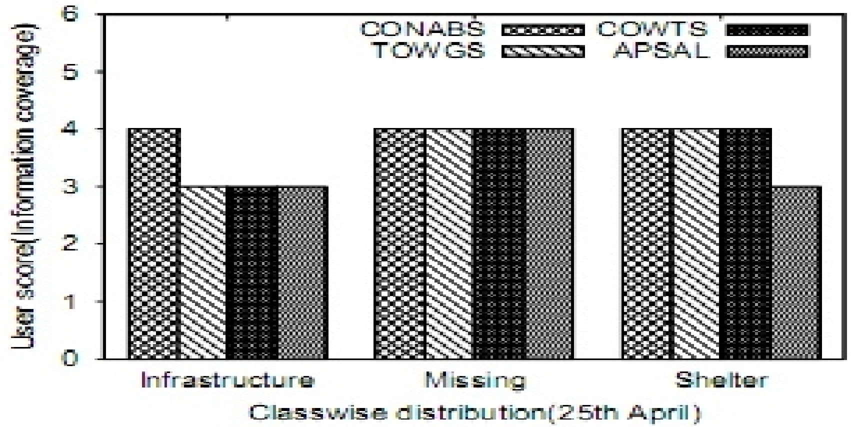
Provide summary for each of the three classes from 25th April to 27th April
Compared against a gold standard summary report generated by experts like **SBTF, UNOCHA**
Generate a system summary of 200 words for each of the three classes across six days

Summarization result

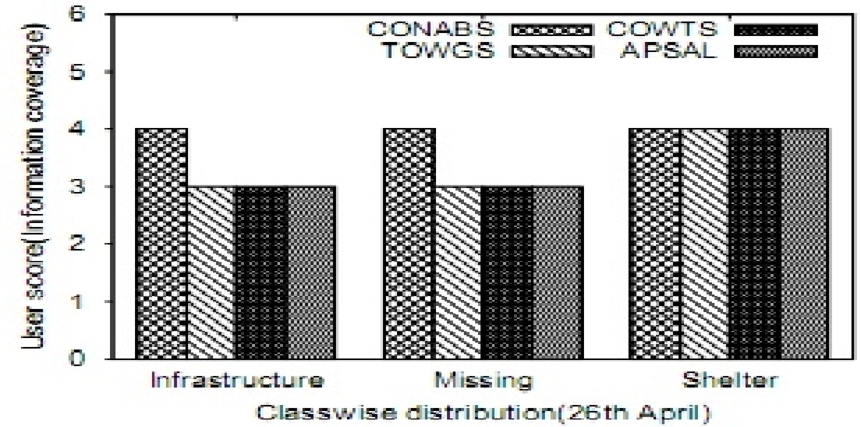


Obtain 20-40% improvement over baselines

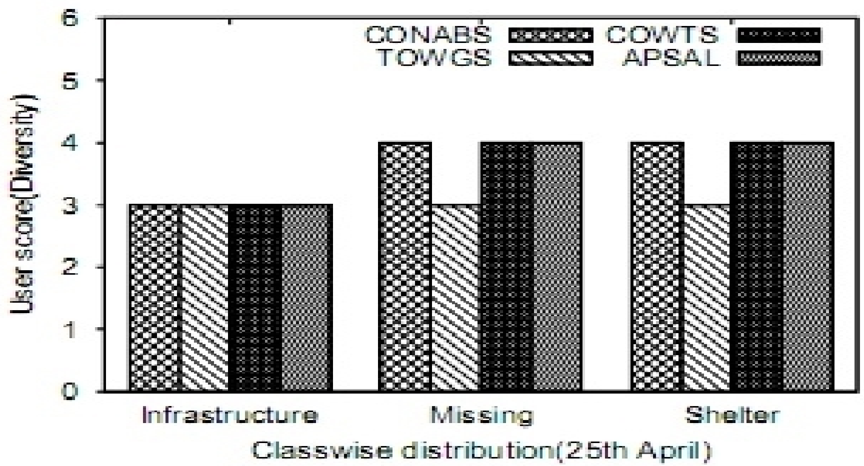
Information coverage and diversity



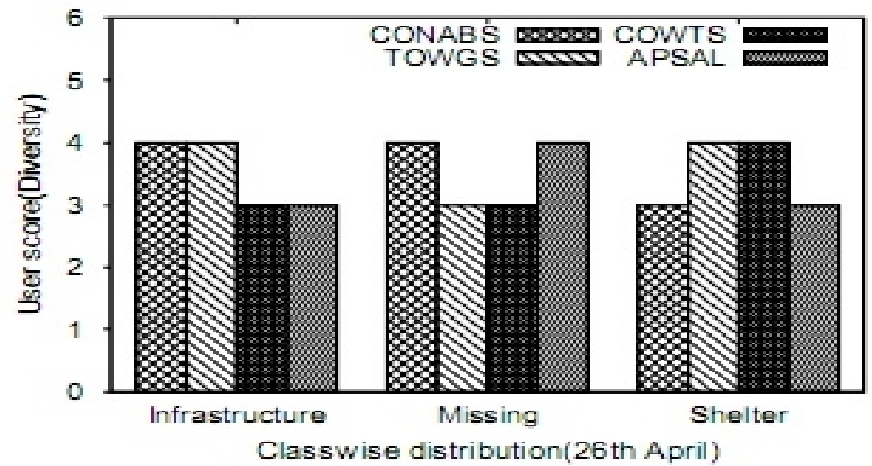
(a) 25th (info. coverage)



(b) 26th (info. coverage)



(a) 25th (info. diversity)



(b) 26th (info. diversity)

Sub-topic Identification

Sub-topic identification

- **Objective:** to capture small-scale sub-events such as ‘power outage’, ‘bridge closure’ etc.
- sub-topic as a combination of a noun and a verb where noun represents a concept and verb represents an event

Topic phrases

Table 3: Popular topic phrases posted on the first day of the Nepal earthquake (Apr 25, 2015)

Class	Topic phrases
Infrastructure	'service affect', 'shut flight', 'crack road', 'water report', 'topple tower'
Injured	'casualty grow', 'victim treat', 'hospital accommodate', 'man trap', 'casualty injure'
Missing	'family stuck', 'tourist strand', 'rescue location', 'database track', 'contact number'
Shelter	'field clean', 'water equip', 'emergency declare', 'deploy transport', 'deploy aircraft'

zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzz zzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzzz

- Consider event words like **killed, injured, died** etc. [Ritter 11]
- Identify nouns directly modify the **events**
  - #China media says **buildings toppled** in #Tibet [url]
  - India **sent** 4 Ton **relief** material, Team of doctors to Nepal
- Obtain a high precision of **0.92** compared to three word window based approach

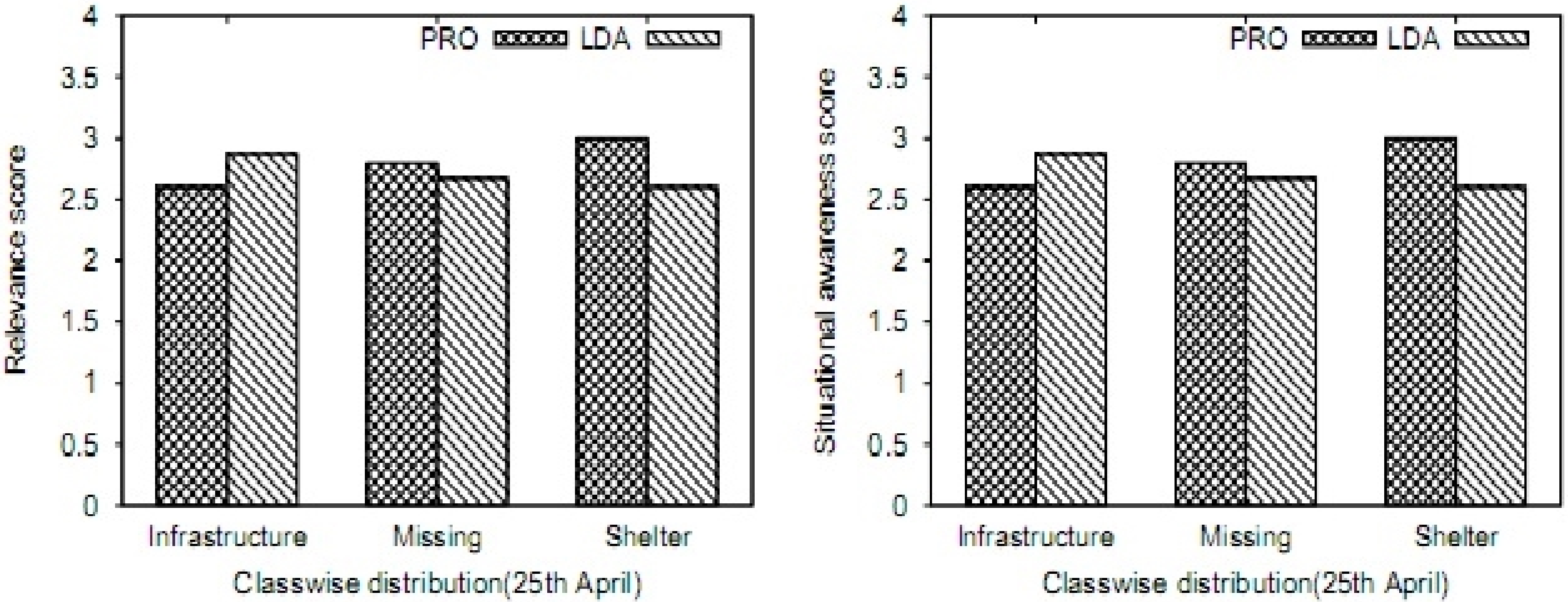
Ranking topic phrases

- Compute Szymkiewicz-Simpson overlap score between noun(N) and verb(E).

$$Overlap(N, E) = \frac{|X \cap Y|}{min(|X|, |Y|)}$$

- X : set of tweets containing N, Y: set of tweets containing E.

Evaluating topic phrases



(a) Relevance (b) Situational awareness

Topic phrases provide relevant as well as important situational information

Alzheimer’s Linguistic Analysis

- Interviews
  - Automatic Coding of Phrases in Interviews
  - Identifying and Suggesting New Codes
  - Modeling Linguistic Variation

Technologies

- Deep Learning
  - Classification Algorithms
    - Text Mining
- Automated Authoring/ChatBots
  - Artificial Intelligence

Questions?

- pmitra@psu.edu