Wherever there are sensations, ideas, emotions, there must be words.

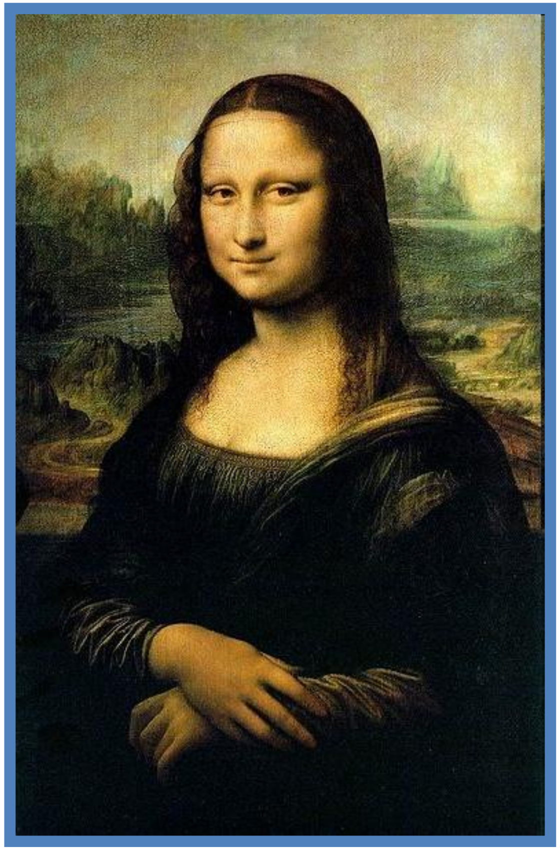
Swami Vivekananda

Acknowledgment: Pushpak Bhattacharyya

This is a talk on 'New Horizons of Sentiment Analysis (SA): SA Research at IIT Bombay'

All images in this presentation are from Wikimedia Commons. Citations missing.

Image source: Wikimedia commons



Mona Lisa

16th Century Portrait

Artist: Leonardo di ser Piero da Vinci

Country of Origin: Florence, Italy

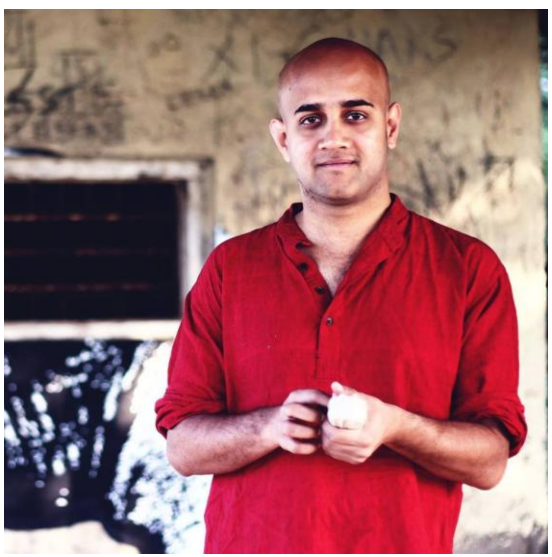
Humans learn all the time

- We learn to drive a car
- We learn to cook Maggi
- We learn to speak

- How does a child learn language?

The story of my year-old nephew

- According to a baby, the first meaning of their name is...?*
- The Kaka anecdote



* : Child language Acquisition; Jean Piaget

How did we learn things?

- The first time we turned a computer on
  - A teacher told me exactly which switch to press, the first time.
  - Later...?
- The first time I came to PICT (in 2012)
  - Google maps or ...?
- The first time my father used Internet banking
  - Step-by-step: one feature at a time

Tell me exactly what to do
Programming

I will do a task based on
Similar tasks in the past: **Learning!**

I will learn each time I do
Something new: **Online learning!**

Sentiment Analysis (SA): Definition

Given a piece of text,

Is she smiling or frowning?

Detect the polarity of opinion expressed by the speaker

Positive?

Negative?

Objective?

I love the movie.

This movie has twenty songs.

This movie is pathetic.

New Horizons of Sentiment Analysis

SA research at IIT Bombay

Aditya Joshi
IIT Bombay | Monash University
adityaj@cse.iitb.ac.in

First presented at DKTE, Ichalkaranji on 10th April, 2015

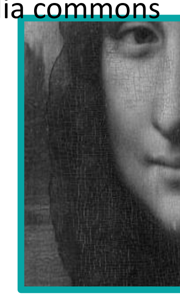
Outline

- Background
- SA for English, Hindi and Marathi
- Emotion analysis for mental health monitoring
- Sarcasm detection

Outline

- Background
- SA for English, Hindi and Marathi
- Emotion analysis for mental health monitoring
- Sarcasm detection

Image source: Wikimedia commons



Challenges

Well-known challenges [8]

- Sarcasm
- Domain Dependence
- Thwarting

Other challenges

- Implicit Polarity
- Interjections/extensions
- Entity Identification

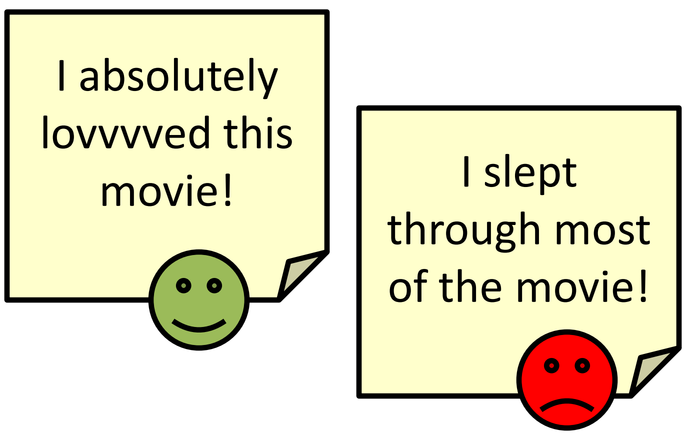
Casablanca and a lunch comprising of rice and fish: a good Sunday

Entity: Casablanca

A basic SA pipeline

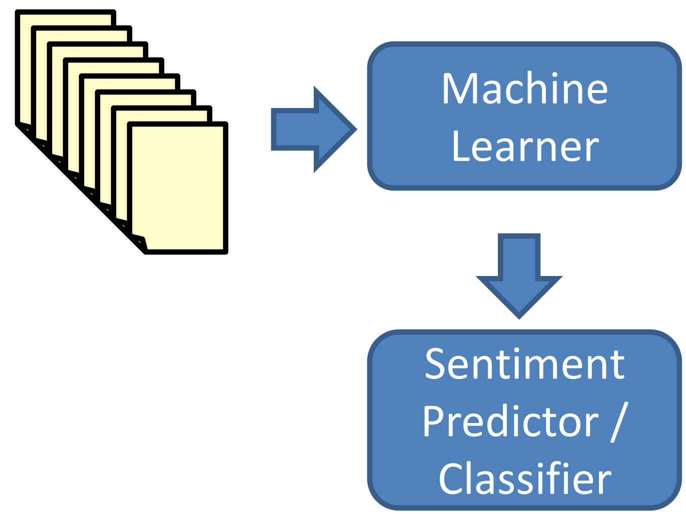
- **What is it?**

Automatic prediction of opinion in text



- **How is it done?**

Use machine learning techniques to learn rules



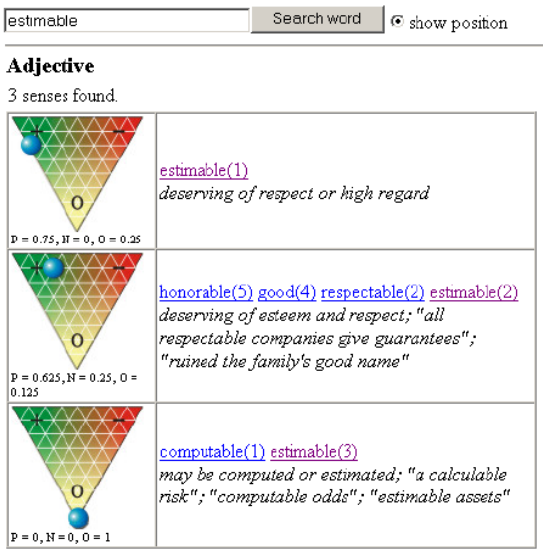
Excellent & rib-tickling -> positive
Boring & predictable -> negative

Outline

- Background
- **SA for English, Hindi and Marathi**
  - **Hindi SentiWordnet**
  - **Using word senses**
- Emotion analysis for mental health monitoring
- Sarcasm detection

Hindi SentiWordnet

- SentiWordnet is a lexical resource that assigns a positive, negative and objective score for synsets in Wordnet



- Originally in English, our Hindi SentiWordnet is the first Hindi adaptation

Creation

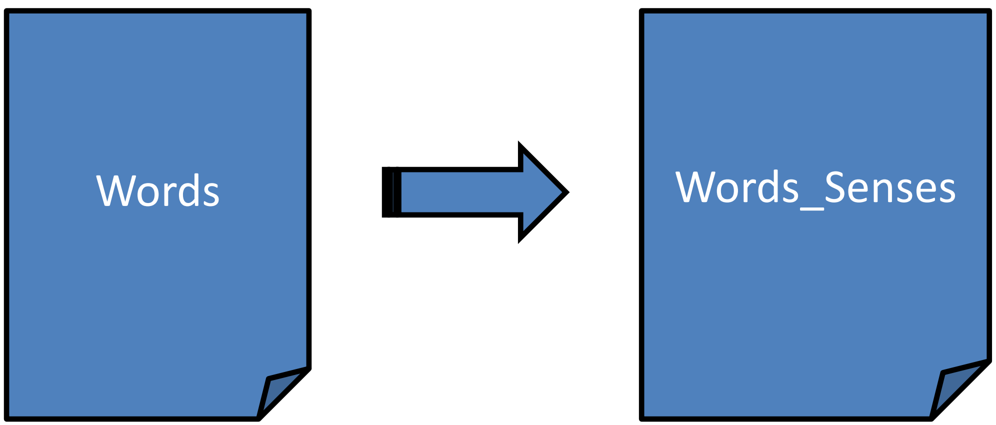
- Creation of Hindi SentiWordNet (H-SWN):
  1. For each synset in the SWN, repeat 2 to 3:
  2. Find corresponding synset in Hindi WordNet using Multidict.
  3. Project the scores in SWN to the synset in Hindi WordNet.
  4. The resultant is H-SWN (16,253 synsets) with sentiment-related scores associated with synsets in the form of triples.

Application

- Using H-SWN for SA:
- For each word in the document,
  1. Apply stop word removal and stemming
  2. Look up the sentiment triple for each word in the H-SWN.
  3. Assign to a word the polarity whose score is the highest.
- Assign to a document the polarity which majority of its words possess.

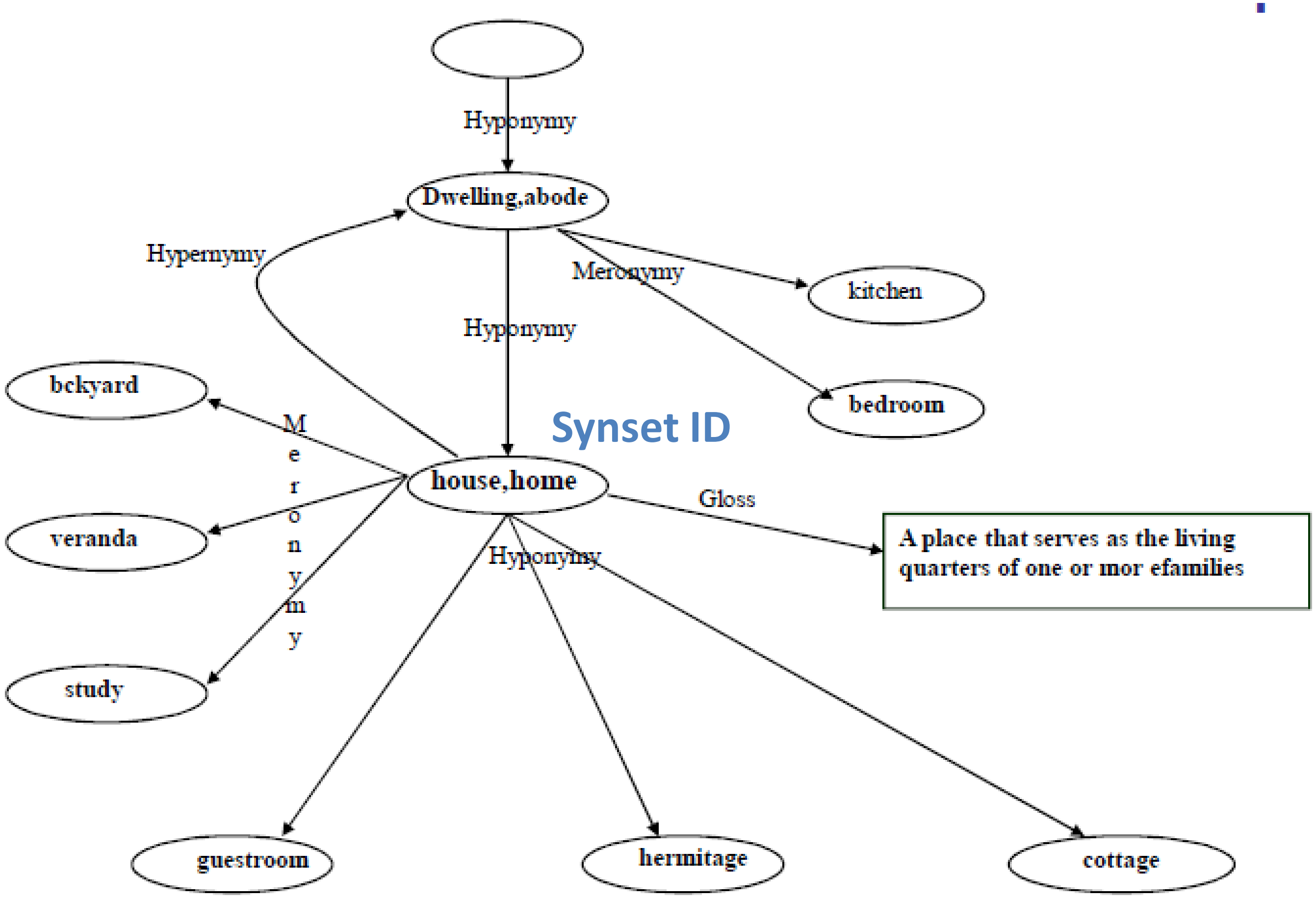
SA of English, Hindi, Marathi **using Word Senses**

Document containing **words** are annotated with **word senses**



Goal: To understand how **word senses** perform as features

Digression: Word senses and Wordnet



Motivation

1. A word may have some sentiment-bearing and some non-sentiment-bearing senses

2. A word may have senses that bear sentiments of opposite polarity

3. A sense can be manifested using different words

Instead of using words as features, we use **Wordnet synset identifiers corresponding to words**

"He speaks a **vulgar** language."
"Now that's real **crude** behavior!"

Lexical space v/s sense space

There are also fire-pits available if you want to have a bonfire with your friends .

Lexical Space

There are also_347757 fire_pits_19147259 available_4203394 if you want_21808093 to have a bonfire_17203241 with your friends_19962226 .

Sense Space

Manually annotated Senses

Automatically annotated Senses

fire_pits : 19147259

(1: POS identifier : Noun, 9147259: Wordnet Synset offset)

Experiment Setup

Data set

- Dataset: Travel Domain Corpus by [33]
- 600 positive and 591 negative travel reviews
- Manual Sense Annotation

Classifier

- C-SVM from Lib-SVM
- Five-fold cross-validation

Results: Overall Classification

Feature Representation	Accuracy	PF	NF	PP	NP	PR	NR
W	84.90	85.07	84.76	84.95	84.92	85.19	84.60
M	89.10	88.22	89.11	91.50	87.07	85.18	91.24
W+S(M)	90.20	89.81	90.43	92.02	88.55	87.71	92.39

- Senses give better overall accuracy
- Negative Recall increases

Using Word Senses for Hindi and Marathi

- **SA of Hindi & Marathi**
  - Lack of resources and classifiers

Sentiment Predictor / Classifier

बेहतरीन (behtareen) (excellent)-> positive
बकवास (bakwaas) (bad) -> negative

अप्रतिम चित्रपट!
(aprateem chitrapat)
(Excellent movie!)

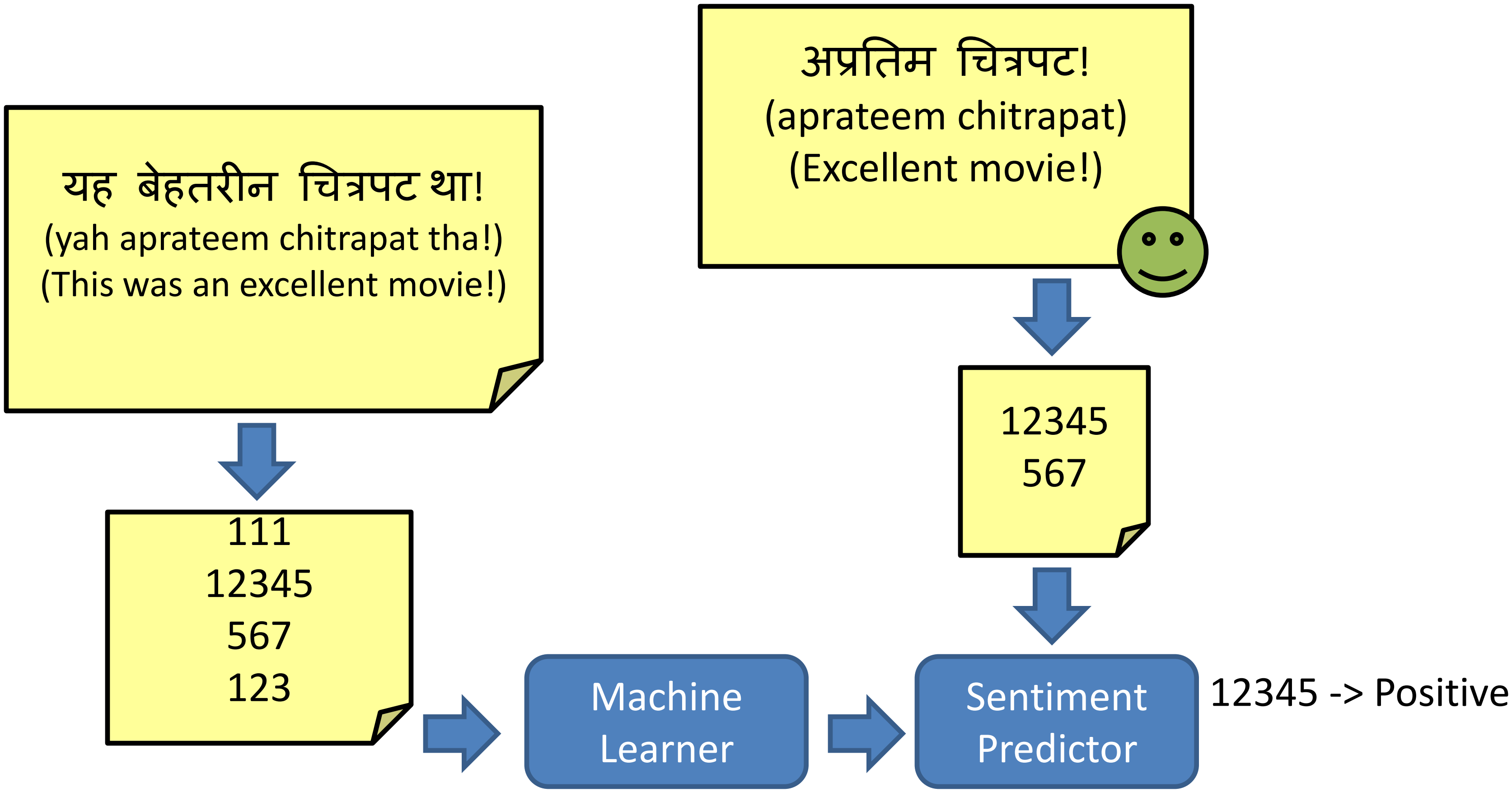
- **How could it be done?**
  - Translation
  - **Cross-lingual SA using Wordnet senses**

Hindi Wordnet	Marathi Wordnet
बेहतरीन (behtareen), उमदा (umda): **12345**	अप्रतिम (aprateem), अजोड (ajoD): **12345**

Dummy entries in Wordnet

How can Hindi and Marathi help each other?

- Learn the sentiment predictor using ‘Wordnet meaning identifiers’ instead of words



Cross-lingual SA Results

- Target Language: Marathi

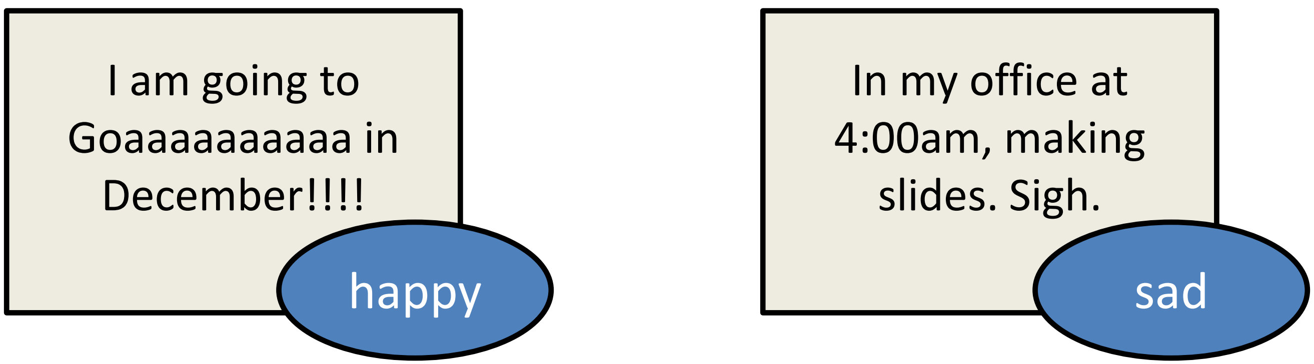
Feature Representation	Accuracy	PF	NF	PP	NP	PR	NR
Translation	71.64	72.22	62.86	75.36	67.69	69.33	58.67
Senses (M)	84.00	81.54	85.88	96.36	76.84	70.67	97.33

Outline

- Background
- SA for English, Hindi and Marathi
- Emotion analysis for mental health monitoring
- Sarcasm detection

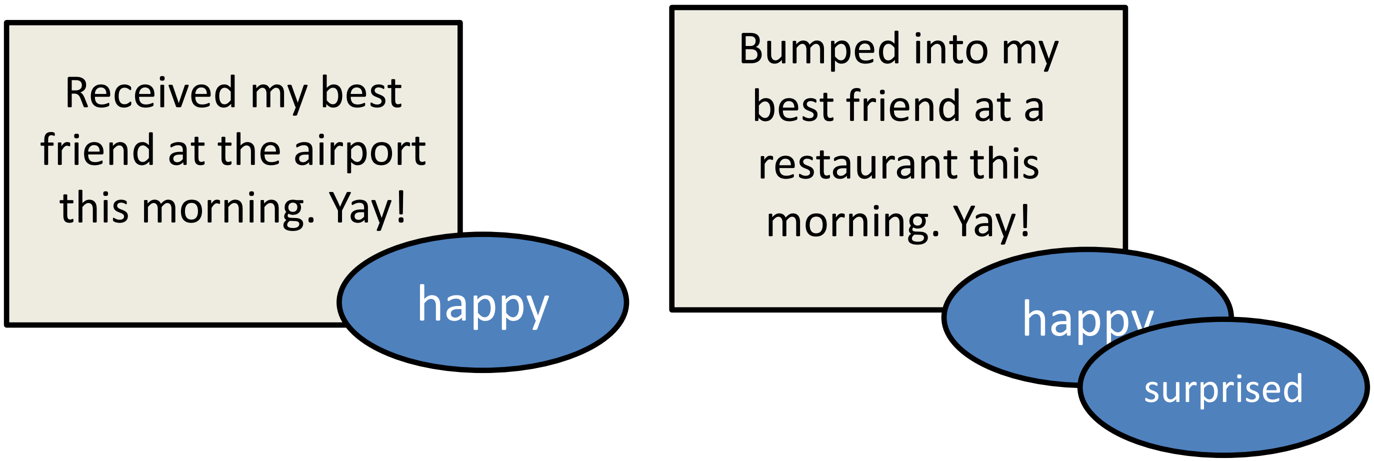
Emotion Analysis (EA): Definition

Emotion analysis of text is defined as the task of predicting emotion expressed in the text



Emotion Analysis and Sentiment Analysis?

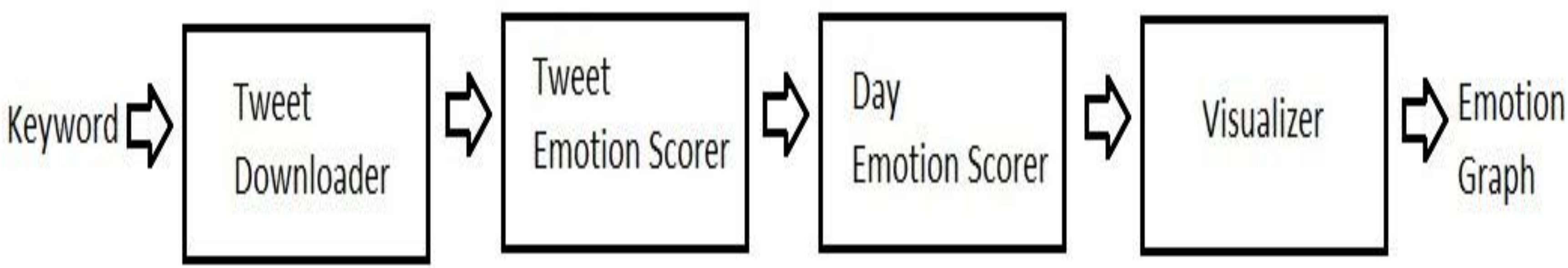
- **Labels: SA** : positive/negative, **EA:** Finer labels like Anger, Disgust, Happiness, etc.
- **Singularity:**



EmoEngine 1.0

- An emotion engine pivoted on the time axis.
- A web-based portal that displays sentiment in tweets on a time-based axis.
- Rule based system which makes use of LIWC emotion lexicon.
- Keywords based search and generate emotions.
- Keyword can be emotion holder or emotion target.
  (Emotions “of” Vs. Emotion “about”)

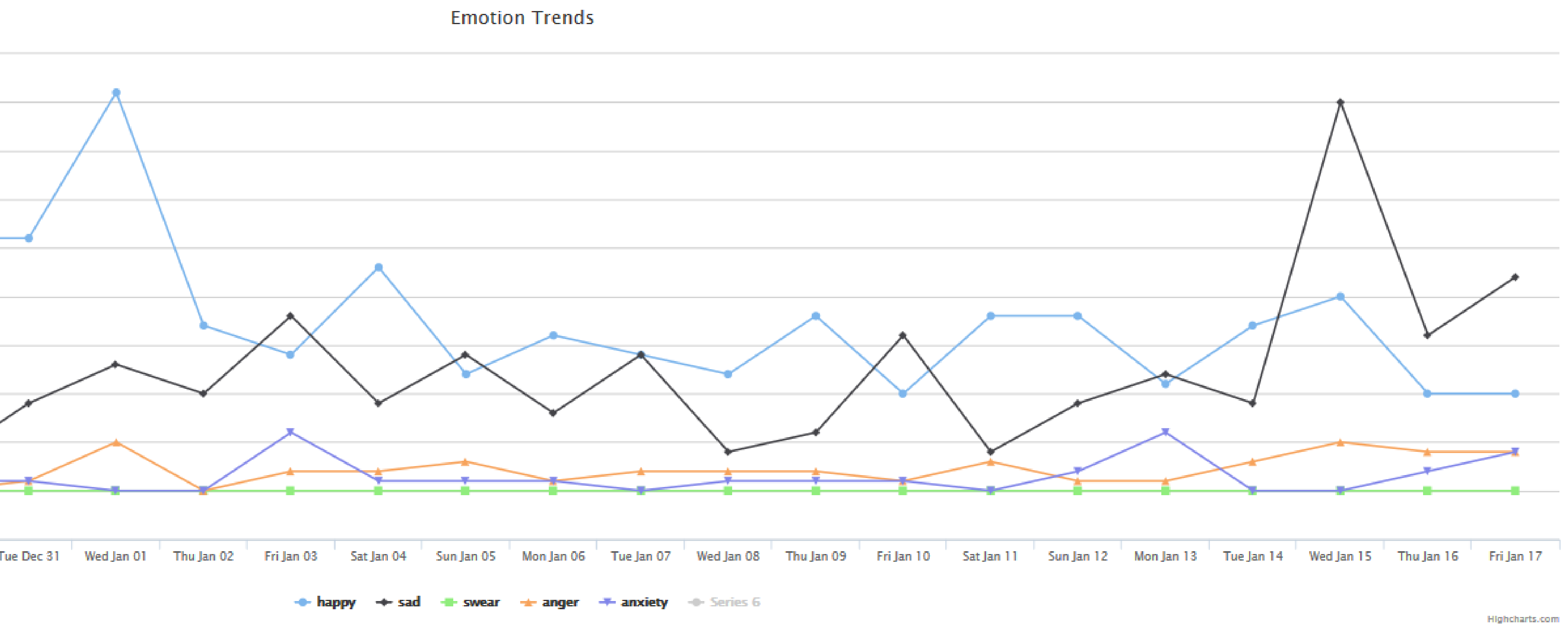
EmoEngine 1.0 : Architecture (1/2)



EmoEngine 1.0 : Architecture (2/2)

- **Twitter Downloader:** Downloads tweets based on the option selected and the keyword
- **Tweet Emotion Scorer:** Assigns an emotion score to each tweet
- **Day Emotion Scorer:** That assigns an overall emotion score to a day
- **Visualizer:** That represents emotion lines on the visual output

Screenshot



Currently hosted at: www.cse.iitb.ac.in/~ravisoni/emotions.htm

EmoEngine: Applications

- Keep track of mental health of person. Useful for her closed ones. (Friends, family, doctor)

- Enhanced version of the engine can effectively be used to understand suicide risks or symptoms of mental health concerns.

- Can be used in advertisement and businesses for customer retention.

Outline

- Background
- SA for English, Hindi and Marathi
- Emotion analysis for mental health monitoring
- **Sarcasm detection**

Sarcasm Detection

- Sarcasm is the presence of words of one polarity in a sentence with another implied polarity.

Example: Being stranded in traffic is the best way to start the week.

- ***Sarcasm is a challenge to SA***
- ***Hypothesis:*** Sarcasm can be identified using a \`sentiment flip’
- We present a sarcasm detection system based on sentiment flip features

Sentiment Flip

Sentiment flip is the inversion in sentiment over the sequence of words in text (tweet in our case)

Implicit Flip

Sentiment words not used. The sentiment is implied.

E.g. I love visiting the dentist twelve times in a month

Implicit Flip features are based on prior work in rule-based sarcasm detection.

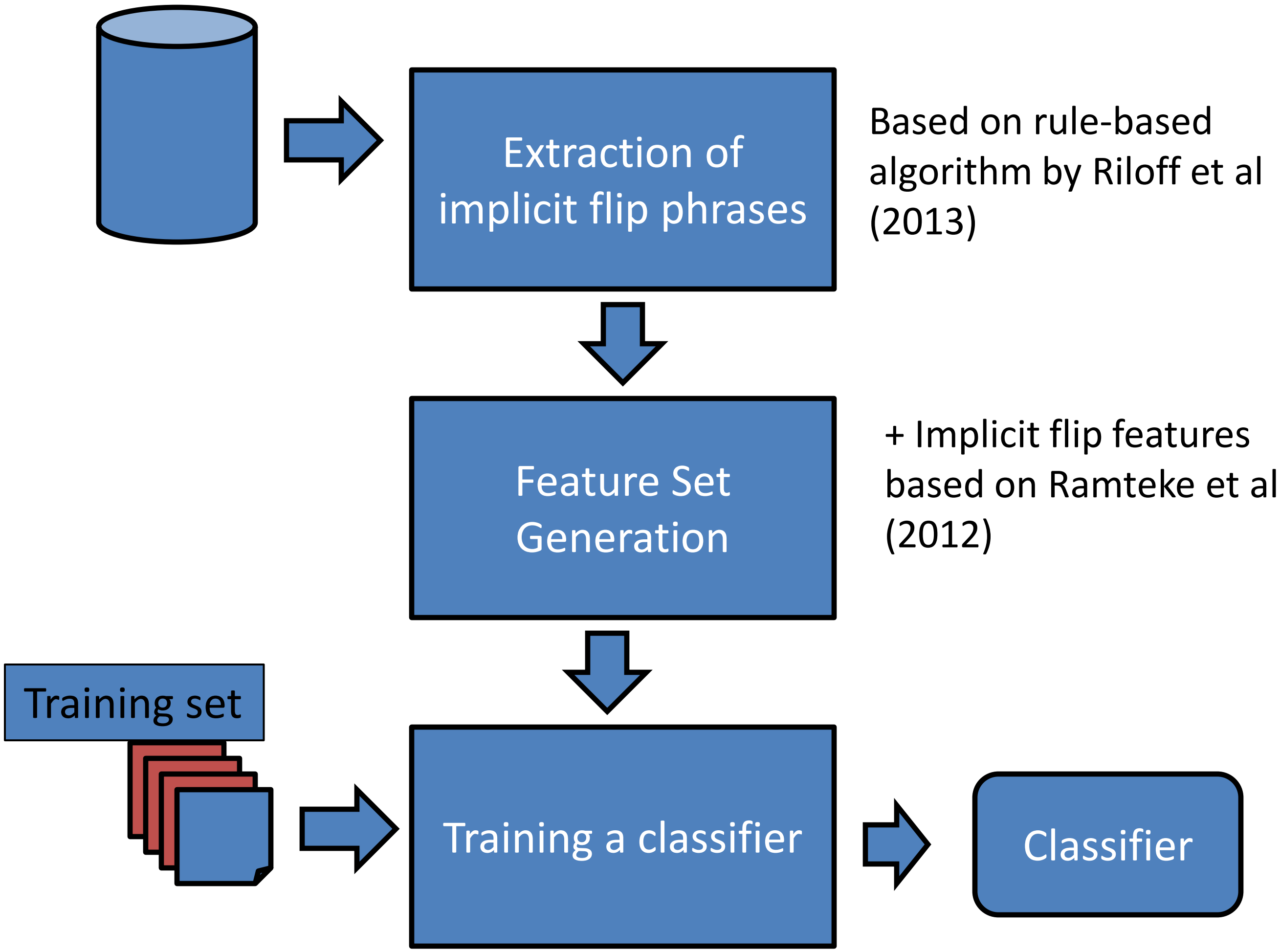
Explicit Flip

Sentiment words of both polarity are used.

E.g. I love being ignored

Explicit flip features are based on prior work in thwarting.

Our Sarcasm Detection System



Implicit flip phrases are two sets of words: (a) positive verbs, and (b) negative situation phrases (e.g. "being ignored")

Datasets

Dataset A (4000 tweets, 50% sarcastic) : To extract implicit flip features

Dataset B (5208 tweets, 4170 sarcastic): Using #sarcasm and related hashtags

Dataset C (15930 tweets, 7218 sarcastic): Using #sarcasm and related hashtags

Dataset D (2278 tweets, 506 sarcastic): Manually annotated by Riloff et al. (2013).

Evaluation

- LibSVM with linear kernel, 5-fold cross-validation

System	Precision
System by Riloff et al. (2013) **(Best value reported)**	0.70
Our system (all features)	**0.772**

Features	P	R	A
Dataset B			
Lexical **(Baseline)**	0.820	0.867	0.741
Lexical+Implicit flip	0.822	0.887	0.755
Lexical+Explicit flip	0.807	0.985	0.801
Lexical+Implicit flip+Explicit flip	0.813	0.978	0.801
Lexical+Implicit flip+Explicit flip+Polarity	0.812	0.977	0.801
Lexical+Implicit flip+Explicit flip+Polarity+ Pragmatic	0.814	0.976	**0.803**
Dataset C			
Lexical+Implicit flip+Explicit flip+Polarity+ Pragmatic	0.767	0.753	0.763

Outline

- Background
- SA for English, Hindi and Marathi
- Emotion analysis for mental health monitoring
- Sarcasm detection

- **Conclusion**

Conclusion

- Traditional approaches to SA use statistical classifiers using unigrams (words), etc.
- We explored use of word senses that enabled SA for Hindi and Marathi
- We developed EmoEngine 1.0: an emotion analysis engine that predicts emotional well-being of a person
- We discussed an approach to sarcasm detection based on sentiment flips

thank you.

http://www.cse.iitb.ac.in/~adityaj

adityaj@cse.iitb.ac.in

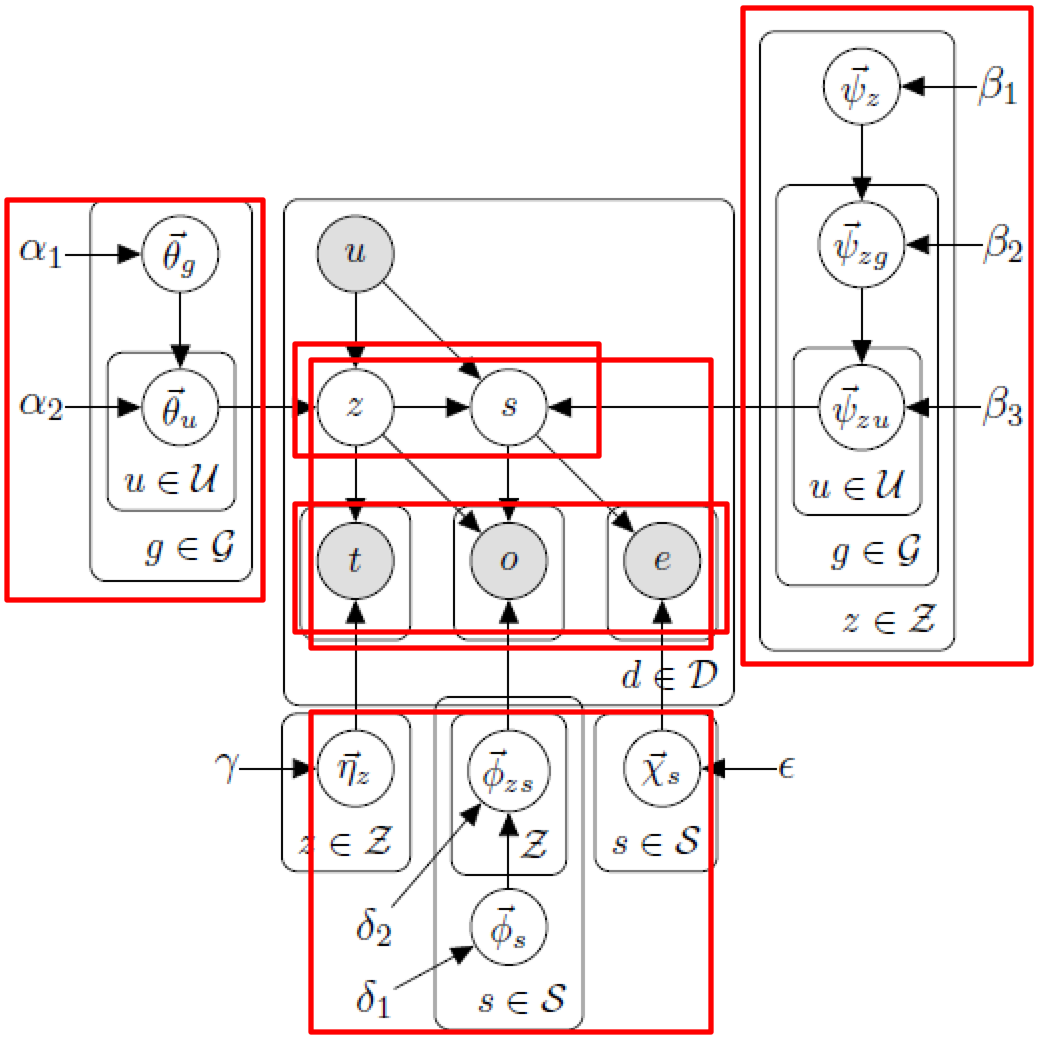
Extra slides

Political Topic Model

- **Goal:** To understand how political issues divide people on Twitter
  - People are divided into “groups”: a group often shares an ideology towards an issue, with some variance among followers of the group
  - We wish to understand: (A) What are the political issues, (B) In what way are groups divided on these issues

- Sentiment-based Political Issue Extraction (SPIE) model

SPIE Model



α_1
α_2
$\vec{\theta}_g$
$\vec{\theta}_u$
$u \in \mathcal{U}$
$g \in \mathcal{G}$
u
z
s
t
o
e
$d \in \mathcal{D}$
$\vec{\psi}_z$
$\vec{\psi}_{zg}$
$\vec{\psi}_{zu}$
β_1
β_2
β_3
$u \in \mathcal{U}$
$g \in \mathcal{G}$
$z \in \mathcal{Z}$
γ
$\vec{\eta}_z$
$z \in \mathcal{Z}$
$\vec{\phi}_{zs}$
$\mathcal{Z}$
$\vec{\phi}_s$
$s \in \mathcal{S}$
δ_2
δ_1
$\vec{\chi}_s$
$s \in \mathcal{S}$
ϵ

Dataset Creation

1) **US Political Tweets**
- 32 Republicans and 46 Democrats
- Expand them by selecting friends
- Complete timeline of users is downloaded
- 24 million tweets

2) **IN Political Tweets**
- 0.5 million tweets

3) **PK Controversy** (In progress)

Political Issues Extracted (1/2)

Insurance	**Abortion**	**Security/War**	**Employment**
health	abortion	attack	workers
insurance	baby	video	job
people	babies	security	wages
care	freedom	police	jobs
plan	women	forces	people
Gun laws	**Immigration**	**Economy**	**Climate**
people	workers	tax	climate
gun	immigration	jobs	people
laws	stories	debt	change
guns	patriot	taxes	warming
control	politics	spending	years
Marriage	**Election**	**Disasters**	**Crime**
people	bill	acres	scene
freedom	vote	fire	police
marriage	campaign	weather	man
rights	state	snow	fire
women	election	storm	suspect

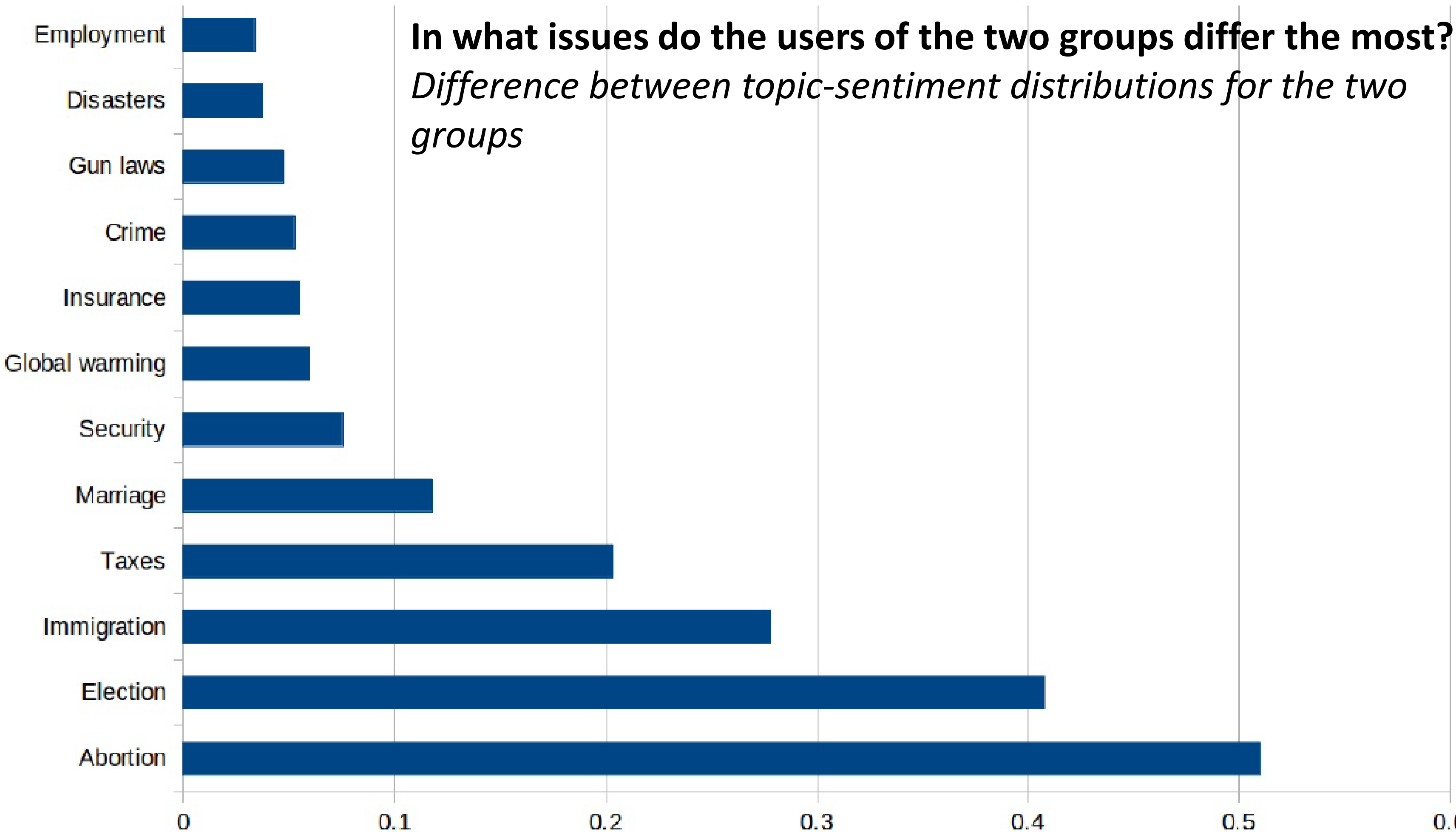
Political Issues Extracted (2/2)

Abortion		Security/War	
Join	Prolife	Killed	Military
Religious	Killed	Syrian	Illegal
Stand	Born	Military	Russian
Support	Unborn	Fast	Targeting
Conservative	Aborted	Furious	Back
Gun laws		**Immigration**	
Illegal	Don	Join	Top
Free	Free	Support	Enter
Don	Stop	Back	Check
Vote	Illegal	Stand	Stop
Stop	Give	Proud	Join
Insurance		**Marriage**	
Pay	Check	Back	Gay
Federal	Hear	Don	Religious
Signed	Here	Lost	Political
Paid	Call	Liberal	Free
Uninsured	Hope	Great	**

Qualitative Evaluation: What are the most “controversial” political issues?

In what issues do the users of the two groups differ the most?

Difference between topic-sentiment distributions for the two groups



Quantitative evaluation: Effect of individual/group distributions

	Average cosine similarity	
	With political affiliation information	**Without political affiliation information**
Within members of the same group		
Democrat-Democrat	**0.261**	0.253
Republican-Republican	**0.014**	0.013
Within members of different groups		
Democrat-Republican	**0.108**	0.113
Republican-Democrat	**0.040**	0.042

Pilot study: To predict political affiliation

	CosineSim(Author, Democrats)	CosineSim(Author, Republicans)
Actual Group Label: Democrats		
U1	**0.772**	0.751
U2	**0.787**	0.727
U3	**0.827**	0.818
U4	0.75	**0.827**
U5	**0.771**	0.764
U6	**0.804**	0.764
Actual Group Label: Republicans		
U7	0.755	**0.818**
U8	0.791	**0.828**
U9	0.803	**0.825**
U10	**0.866**	0.84
U11	0.763	**0.78**
U12	**0.813**	0.812
U13	**0.835**	0.788
U14	0.842	**0.843**
U15	**0.839**	0.782
U16	**0.826**	0.795
U17	**0.876**	0.82
U18	**0.736**	0.713
U19	0.794	**0.805**
U20	0.744	**0.755**
U21	**0.769**	0.715
U22	**0.824**	0.798
U23	0.808	**0.811**
U24	0.8	**0.85**
U25	0.774	**0.785**

Political Topic Model: Summary

- We implemented SPIE model and tested it on multiple datasets
- Our qualitative and quantitative evaluation shows that the model performs well
- We also ran a pilot experiment to see how this model can be used to predict political orientation