Characterizing Microblogs with Topic Models

ICWSM 2010

Daniel Ramage
Stanford NLP

Susan Dumais, Dan Liebling
Microsoft Research
Microblogging on Twitter

15 minutes tune in to my Oscar special. Some cool celebs interviewing each other.
6:51 PM Mar 3rd via web

RT @SheriSalata great show with Nate today. anniv. of tsunami. dealing with grief. @TheOprahShow
6:16 AM Jan 14th via web

#FF @theoprahshow - lots of show info here
2:51 PM Jan 8th via web

Hanging out with friends - "pom" martinis-getting ready to watch xmas special. 10 eastern 9 central. Going caroling afterward!
6:43 PM Dec 13th, 2009 via web

140 characters or less  Refer or reply to other users  Inline tagging with hashtags  Following ≠ friending
Do people like the posts they see?

43 users at Microsoft judged 60 or more posts
Do people like the posts they see?

“not really worth reading”
“maybe worth the time spent reading”
“worth the time spent reading”
Do people like the posts they see?

Number of judges with average rating

“not really worth reading”  =
“maybe worth the time spent reading”  =
“worth the time spent reading”  =
The fundamental problem

People followed $\neq$ Tweets worth reading
What factors go into following decisions?

Structured interviews with heavy users (at Microsoft)
Broader survey of 56 Twitter users (at Microsoft)

**Substance**
- Hobbies, profession, news, products, events
- Humor, wit, whininess, diction, worldview

**Status**
- Making plans, networking, staying in touch
- Updates about meals, travel, hygiene

**Social**

**Style**
- Humor, wit, whininess, diction, worldview
Content modeling on Twitter

8.2m “Spritzer” posts from November 2009
About 13 words each, after filtering
Content modeling on Twitter

Surface word features

tf.idf cosine similarity, etc.

Deeper natural language processing

Parsing, parts of speech, coreference, etc.

dats yur mom not me lol
THE_REAL_SHAQ
Content modeling on Twitter

Surface word features

Topic models, Dimensionality reduction

Supervised classification

#hashtags, emoticons, questions, etc.

Naïve Bayes, SVM, etc.

Best model in ranking experiments

Labeled LDA

tf.idf cosine similarity, etc.

Latent Dirichlet Allocation (LDA), LSA, etc.
Content modeling with Labeled LDA

Discover unlabeled topics
Parameter K=200 latent topic dimensions

Model common labels
500 - 1000 dimensions for hashtags, emoticons, etc.

- obama
- president
- american
- america
- says
- country
- russia
- pope
- island

- I’m going go out
- gonna see im
- tonight sleep
- tomorrow about
- am night

- Smile : )
  - :) good day
  - morning thanks
  - have happy
  - hope birthday

- :) can’t wait see
- one yay!!! cant
- tomorrow got !!
- next christmas

#jobs
- #jobs featured
- manager sales
- engineer yahoo
- location senior
Content modeling with Labeled LDA

new muppetblog political commentary link

@kermit heyy wanna catch a movie

just ate a cookie #yummy

Histogram as signature for set of posts
Twitter content by category

- Aggregate topics into (one or more) 4S categories: substance, status, social, style, or other
- Good inter-annotator agreement for latent topics. Fleiss’ $\kappa$ between 0.754 substance to 0.370 social
- Shortcut: hashtags are substance, replies are social, emoticons are social and style
can make help if someone tell_me them anyone use makes any sense trying explain

up what's hit pick whats hey set twitter sign give catch when show first wats make

haha lol :) funny :p omg hahaha yeah too yes thats ha wow cool lmao though kinda

im get dont gonna shit gotta wanna cuz damn ur make cant say cause bout ill mad tired

obama president american america says country russia pope island failed honduras

iphone new phone app mobile apple ipod blackberry touch pro store apps free android an

am still doing sleep so going tired bed awake supposed hell asleep early sleeping sleepy

night sleep bed going off tomorrow bye tonight goodnight all im time now nite
Characterizing Microblogs with Topic Models

Outline

• Modeling Twitter content with topic models
• Characterizing, recommending and filtering
Characterizing users
Characterizing users
Characterizing Microblogs with Topic Models

Outline

• Modeling Twitter content with topic models
• Characterizing, recommending and filtering
Ranking experiments on filtering & finding

Twitter Feed Rater

steephill @rkrzyston I can see Deadwood becoming a cult classic. Great writing and acting. Already thought about watching it again. Nov 10 2009 7:14PM

timoreilly Note especially education diffs. RT @ahier: Jobless Rate for People Like You - fascinating interactive graphic http://bit.ly/4b3szK Nov 10 2009 6:51PM
Filtering: Tweet stream re-ranking

Split rater’s posts into train (70%) and test (30%)
Re-rank test set by distance to positive examples
Consider judgment of □ or ↑ as “positive”
Filtering: Tweet stream re-ranking

Split rater’s posts into train (70%) and test (30%)
Re-rank test set by distance to positive examples
Consider judgment of □ or ▶ as “positive”

Feature space:

- Cosine similarity tf.idf
- Labeled-LDA topic space
- Weighted combination

Surface word features
Dimensionality reduction + classification
Filtering: Tweet stream re-ranking

Split rater’s posts into train (70%) and test (30%)

Re-rank test set by distance to positive examples

Consider judgment of ▼ or ▶ as “positive”

Mean Reciprocal Rank @ 1 Relevant
Finding: User recommendation task

Rater’s followed users: train (6/7) and test (1/7)
Find the test user among 8 other non-followed users

Ranking task: score by reciprocal rank of test user
Characterizing Microblogs with Topic Models

• Content analysis on microblogs can help unmet needs: characterizing, recommending, and filtering
• Labeled LDA discovers topics that align with 4S categories, provides insight into language use; useful for ranking
• Next steps: better interfaces for finding and filtering and models that account for temporal dynamics

http://twahpic.cloudapp.net/

Characterizing Microblogs with Topic Models
Daniel Ramage, Susan Dumais, and Dan Liebling
Backup Material
Learning and Inference

- Need to scale up to Twitter-size datasets
- Strategy: move from Gibbs sampling (inherently sequential) to variational inference (optimization based)
- Recipe: save sampling distributions for each topic assignment; batch updates (Asuncion et al UAI 2009)

Assignments

- Sequential
- Batch

Update strategy

- Collapsed Gibbs Sampler
- CVB0

Assign

\[
\gamma_{d,i,k} \propto \frac{\gamma_{dik} + \eta}{\gamma_{dik} + \eta_d} \cdot (#d_k - \gamma_{dik} + \alpha) \cdot I[k \in \Lambda_d]
\]

Count

\[
\begin{align*}
#d_k &= \sum_i \gamma_{d,i,k} \\
#kw &= \sum_i \gamma_{d,i,k} \cdot I[w_{d,i} = w] \\
#k &= \sum_w #kw
\end{align*}
\]
Filtering: Tweet stream re-ranking
Do people like the posts they see? No.

1. Not really worth reading
2. Maybe worth the time spent reading
3. Must read
Twitter Data

- 8.2m posts from Twitter’s “spritzer” feed, collected November 17th – 24th 2009
- Tokenization honors URLs, emoticons, hashtags, multi-word entities (e.g. michael_jackson)
- Average tweet length of 13.1 words from a vocabulary of 5,119,312 words
Topic categories

- **Substance** topics
  #hashtags, some latent topics
- **Social** topics
  emoticons, replies, @user, questions
- **Status** topics (latent)
- **Style** topics (latent)

<table>
<thead>
<tr>
<th>Category</th>
<th># Topics</th>
<th>Fleiss’ $\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substance</td>
<td>35/127</td>
<td>0.747</td>
</tr>
<tr>
<td>Status</td>
<td>19/127</td>
<td>0.598</td>
</tr>
<tr>
<td>Style</td>
<td>41/127</td>
<td>0.563</td>
</tr>
<tr>
<td>Social</td>
<td>11/127</td>
<td>0.370</td>
</tr>
<tr>
<td>Other</td>
<td>34/127</td>
<td>0.833</td>
</tr>
</tbody>
</table>
Characterizing users

@oprah

@oprah's topic usage [a]:

Latent Topic Space

@w3c's topic usage [b]:

W3C

Name W3C Team
Location NTT | ERCIM | Keio University
Web http://www.w3.org/

Best web standards. Make the Web open, free and on everyone's terms.

Great,

web

developers

designers

we need your help.

Welcome to the W3C.
Characterizing users, take two
Approach: Partially supervised topic model

- LabeledLDA [Ramage et al 2009] assumes every word comes from some multinomial distribution (topic).
- Each tweet has its own weighted topic mixture.
- Some topics are constrained to align with document labels (supervised). Other topics are unconstrained and latent (unsupervised).
Realtime characterization

• Live demo at Microsoft TechFest next week, built by Dan Liebling

• Characterize sets of tweets by topics:
  – Tweets by a user
  – Tweets by a set of users (e.g. users followed by @rion)
  – Tweets matching a query

• ~5 seconds to characterize 200 tweets

• ~4 days to train a model on one week’s data
Next steps

• Applications and interfaces to help people get what they want to read from social streams
• Model readers’ preference by 4S category?
• Better interfaces for consuming Twitter content?
• Models that coherently account for the changing data stream over time. Account for varying numbers of sub-topics.
• How about that social graph?
**Characterizing Microblogs with Topic Models**

Daniel Ramage with Susan Dumais and Dan Liebling

---

**Naive Bayes**

- good day morning thanks have happy hope today all thank happy birthday great had made luck smile wish
- night hahaha goodnight two sleep now back site bye x dx
- wait can't see one yay!!! cant tomorrow got!! next christmas here then till finally break home more
- well much!! thank oh very nice x better soon hope amazing was really cool its get again feel
- love your follow hah ha all thanks new back followers ur welcome nice hey hello pic like im shout out thank
- ha cool da french az oi super ben ne ama o mi soo buir ds hehe @parishilton nem ve
- ich und das ist die auch du ja nicht der mal mit dann noch aber mir wieder es ein hab
- I love have im good be but your are was hah a al it 2!! thanks like day not can too up o a new one go we out will im & if know time got what * see its eu from * well follow as how ya la going hope about today * ... back some am oh more much its great better thank happy :D when twitter think please home he really en 1 en want then da y here don't make tomorrow morning.

---

**Unlabeled Trends**

- love song music lady gaga by new album listening songs makes listen amazing bad its cd much her all &
- by man who * was being after police found old dead man sex an young woman sia say killed under
- news us obama has says real us press report as million china state presid 5 india market its deal
- food some eat dinner than -kgiving eating & turkey coffee chocolate hot lunch chicken drink wine cheese
- get go back home work then some got time off from now getting ready need school done today
- good too i'm really not very but bad much well like today work some pretty feel way am doing fun
- I'm oh im am .... fuck !!! shit its hate hell damn here yeah bored fucking what .... tired
- bad go re why think guys fuck okay can't aww gonna broke class has didn't did thats hey larm
- im school not bored miss though too tired yeah 3 & homework being bit hurt only cant either some home
- hope now isn't english says sad ah idk still hurts giving feel threat mad actually tomorrow site aw all
- nao meu è nem com o um mas isso twitter triste hoje em estou tâ vê mm pq sei só
- eu minha sem ah msn quero te mt mae cara da son enquiry internet pc entra ninguém tava sair aaah
- eu o have não im e but not é be like pra im was get up now what mais go mas com 2 all me da nao se don't um na you are too if know want out !! ' as its am really good nem one some em hah :D vai school la today a about & think got x por from back y time tenho see mina can uma need day * still much feel will twitter going bad 3 muto more its per 9 hoje your why ta how
- like ugh x sorry i'm have want hate it's sucks really damn will life hahaha same sick phone seem
- eu o have não im e but not é be like pra im was get up now what mais go mas com 2 all me da nao se don't um na you are too if know want out !! ' as its am really good nem one some em hah :D vai school la today a about & think got x por from back y time tenho see mina can uma need day * still much feel will twitter going bad 3 muto more its per 9 hoje your why ta how

---

**Hashtags**

- #musicmonday by song ft 2 & feat remix music listening listen album @lights love free chris brown go money rihanna
- #music by song ft love & feat usher trey_song5 tywayne baby drake -- remix nicki beyonce off aka j club
- #blogtalkradio listen air show will now come radio live anytime minutes listening by call or interview talk 5 am join
- #n frightening like being good having ur sex some getting your !!! seeing fresh man gettin .... home after real ass own
- #jobs job featured manager sales engineer yahoo location senior hotjobs london client developer business project #value creation consultant
- #food food some eat dinner than -kgiving eating & turkey coffee chocolate hot lunch chicken drink wine cheese
- #unlabeled trend the song music lady gaga by new album listening songs makes listen amazing bad its cd much her all &
- #tuesday thanksgiving turkey thankful thanksgiving & our please give how thanks traveling tips service what if read road today enjoy here's
Filtering: Tweet stream re-ranking

Split rater’s posts into train (70%) and test (30%)
Consider judgment of 2 or 3 as “positive”
Re-rank test set by distance to positive centroid:
- cosine similarity tf.idf
- Labeled-LDA topic space
- Weighted combination

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean Average Precision</th>
<th>Mean Prec@1</th>
<th>Mean RR@1R</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-LDA + tf-idf (best)</td>
<td>0.622</td>
<td>0.634</td>
<td>0.756</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.604</td>
<td>0.537</td>
<td>0.681</td>
</tr>
<tr>
<td>tf-idf</td>
<td>0.608</td>
<td>0.585</td>
<td>0.718</td>
</tr>
<tr>
<td>Temporal</td>
<td>0.565</td>
<td>0.537</td>
<td>0.678</td>
</tr>
<tr>
<td>Random</td>
<td>0.542</td>
<td>0.537</td>
<td>0.670</td>
</tr>
</tbody>
</table>

Surface word features
Dimensionality reduction + classification
Finding: User recommendation task

Rater’s followed users: train (6/7) and test (1/7)
Find the test user among 8 other non-followed users

Ranking task: score by reciprocal rank of test user

<table>
<thead>
<tr>
<th>Model</th>
<th>Reciprocal Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-LDA + tf-idf (best)</td>
<td>0.965</td>
</tr>
<tr>
<td>L-LDA</td>
<td>0.579</td>
</tr>
<tr>
<td>tf-idf</td>
<td>0.839</td>
</tr>
<tr>
<td>Random</td>
<td>0.313</td>
</tr>
</tbody>
</table>
Labeled LDA assumes each tweet’s words generated by some topic’s word distribution.

Topics are either *latent* (used on all tweets) or *labeled* (only used on tweets containing a label).

Labels come from features extracted from tweet or user (emoticons, questions, replies, hashtags, wefollow tags).
### What we learn

#### Hashtags

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Example Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>#musicmonday</td>
<td>by song ft. feat remix music listening listen album @lights love free chris brown go money rihanna #mm</td>
</tr>
<tr>
<td>#mm</td>
<td>by song ft. love feat usher Trey_songz lil_wayne baby drake -- remix nicki beyonce off aka j club</td>
</tr>
<tr>
<td>#blogtalkradio</td>
<td>listen air show will now come radio live anytime minutes listening by call or interview talk 5 am join</td>
</tr>
<tr>
<td>#aintnothinglike</td>
<td>being good having ur sex some getting your !!! seeing fresh man gettin .... home after real ass own</td>
</tr>
<tr>
<td>#jobs</td>
<td>job featured manager sales engineer yahoo location senior hotjobs london client developer business project #vacature consultant</td>
</tr>
<tr>
<td>#tcot</td>
<td>obama not #ucot #sgp #palin obamacare vote palin #glenbeck acorn obama's sarah palin us trial his climategate #obama poll</td>
</tr>
<tr>
<td>#thanksgiving</td>
<td>turkey thankful thanksgiving &amp; our please give how thanks traveling tips service what if read road day enjoy here's</td>
</tr>
</tbody>
</table>

#### Unlabeled Trends

<table>
<thead>
<tr>
<th>Trend</th>
<th>Example Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>love song music lady gaga by new album listening songs makes listen amazing bad its cd much her all &amp;</td>
<td>get go back home work then some got time off from now getting ready need school done today</td>
</tr>
<tr>
<td>by man who ~ was being after police found old death men sex an young woman via say killed under</td>
<td>i'm going go out gonna see im tonight sleep tomorrow about am night bed again ok wanna think soon home</td>
</tr>
<tr>
<td>news us obama has says real u.s. press report as million china state president 5 india market its deal</td>
<td>good too i'm really not very but bad much well like today work some pretty feel way am doing fun</td>
</tr>
<tr>
<td>food some eat dinner thanksgiving eating &amp; turkey coffee chocolate hot lunch chicken drink wine cheese</td>
<td>..... i'm oh im am ..... fuck !!! shit its hate hell damn here yeah bored fucking what ...... tired</td>
</tr>
<tr>
<td></td>
<td>be will would if should could there have ill make might wish not think must may sure going gonna soon</td>
</tr>
</tbody>
</table>

#### Smile : )

<table>
<thead>
<tr>
<th>Comment</th>
<th>Example Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>good morning thanks have happy hope today all thank happy birthday great had made luck smile wish</td>
<td>night haha good love time goodnight tweet twitter sleep now bed off ok sweet nite bye x dreams gotta</td>
</tr>
<tr>
<td>wait can’t see one yay !!! cant tomorrow got !! next christmas here then till finally break home more</td>
<td>i'm excited am im happy be !! yes thanks glad &amp; it’s yay awesome are going getting now tomorrow too</td>
</tr>
<tr>
<td>well much !! thank oh very nice x better soon hope amazing was really cool its get again feel</td>
<td>love your follow haha hi all thanks new back followers ur welcome nice hey hello pic like im shoutout thank</td>
</tr>
</tbody>
</table>

#### Frown : ( |

<table>
<thead>
<tr>
<th>Comment</th>
<th>Example Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad feel today day feeling hurts cold ill had all pain isn’t phone bored working hurt died hard yesterday</td>
<td>have but i’m sick still go school ugh don’t wanna why tomorrow home won’t didn't headache dont get</td>
</tr>
<tr>
<td>not work oh really well back again hope off never going 7 stupid come down without doesn’t soon okay</td>
<td>miss sad too im !!! awww much him baby gonna happy_birthday crying miley makes see nighthawk</td>
</tr>
<tr>
<td>hate wish .... was d x could missing stop help rain ohh noooo week away ugh concert here tear</td>
<td>sorry now can’t cant sleep need am missed find im ok omg didnt suck sore rip because fail about</td>
</tr>
</tbody>
</table>
Microblogging on Twitter

140 characters or less
Refer or reply to other users
Simple inline tweet tagging

15 minutes tune in to my Oscar special. Some cool celebs interviewing each other.
6:51 PM Mar 3rd via web

RT @SheriSalata great show with Nate today. anniv. of tsunami. dealing with grief @TheOprahShow
6:16 AM Jan 14th via web

#FF @theoprahshow - lots of show info here
2:51 PM Jan 8th via web

Hanging out with friends - "pom" martinis-getting ready to watch xmas special. 10 eastern 9 central. Going caroling afterward!
6:43 PM Dec 13th, 2009 via web

Echo others by retweeting
Following ≠ friending
Big celebrity phenomena
Microblogging on Twitter

#FF @theoprahshow - lots of show info here
2:51 PM Jan 8th via web

Hanging out with friends - "pom" martinis-getting ready to watch xmas special. 10 eastern 9 central. Going caroling afterward!
6:43 PM Dec 13th, 2009 via web

THE_REAL_SHAQ
The fundamental problem

People followed \neq Tweets worth reading

Who should I follow?
Characterizing Microblogs with Topic Models

Outline

• Who to follow, what to read on Twitter
• **Modeling Twitter content with topic models**
• Characterizing, recommending and filtering
Characterizing users
Content modeling with Labeled LDA

new muppetblog political commentary link

@kermit hey wanna catch a movie

just ate a cookie #yummy #yummy
Content modeling with Labeled LDA

new muppetblog political commentary link

@kermit heyy wanna catch a movie

just ate a cookie #yummy #yummy
Content modeling with Labeled LDA

new muppetblog political commentary link

@kermit heyy wanna catch a movie

just ate a cookie #yummy
Characterizing users

Oprah  Think you can host YOUR OWN SHOW? Or know someone who can. The search is on. Go to oprah.com for details.
        7:11 PM May 19

Oprah  Great way to spend a Sunday - “Life” marathon all day on Discovery Channel
        4:34 PM April 16

Oprah  Time flies – O Mag celebrates its 10th anniversary May 7-9. Join me and my friends in NYC! Tix @ Oprah.com/Oturns10
        3:51 PM April 05

Oprah  A big No Phone Zone shout-out to ABC’S Cougartown! That’s right guys, talking on your phone while driving IS crazy-dangerous
        1:58 PM March 26

Oprah  15 minutes tune in to my Oscar special. Some cool celebs interviewing each other.
        6:51 PM March 03

Oprah  RT @SheriSalata if you need some perspective today about life watch film critic roger ebert on @TheOprahShow today. it is an uplifting WOW.
        7:51 AM March 02

Oprah  RT @sherisalata drew brees on today with his gorgeous wife and little baby boy—big who dat fun.
        11:11 AM February 12

Oprah  RT @sherisalata big moment with the fabulous Celine Dion today on @theoprahshow. surprise surprise surprise
        6:41 AM February 10