Social Media Analytics
Text Mining for User Profiles and More

Sofus A. Macskassy
Data Scientist, Facebook
Project Leader, Research Assistant Professor, USC/ISI (on leave)
(sofmac@gmail.com)
Why Mining Social Media?

1. It is the pulse of the world
   - People
   - Events
   - News
   - Real-time

2. Rich and complex
   - Micro-blogs
   - Blogs
   - News
   - Videos
   - Pictures
   - ...

3. Techniques broadly applicable
Making sense of social media

• **What**...
  • Is going on in the world
  • Are people taking about

• **Where**...
  • Is X happening
  • Are people talking about X

• **When**...
  • Did X happen
  • Was X talked about

• **Who**...
  • Is talking about X
  • Is influencing the discussion about X

• **Dynamics**
  • How are things spreading and changing?
Generality of Principles and Techniques

• Complexity of dynamics in content and network
  • Covers a wide variety of disciplines including machine learning, network science, physical systems, social science, economics, game theory, epidemiology, etc.

• Applicable to many problems
  • Better understanding and modeling of social systems
  • Network analysis in financial markets, as defined by social networks, money transactions, etc.
  • Tracking information diffusion by demographics, geospatial, topics, etc.
  • Deep analysis of group-behaviors over time
  • Reasoning with uncertainty about observations and nodes in a network
  • ...

Technical Challenges: Many tools still nascent

1. Large and growing amount of data
   - Many sources
   - Noisy and redundant
   - Streaming

2. Data is complex, varied and changing
   - Entity types are increasingly complex
   - Social network keeps changing
   - Resolving topics, entities, etc. across streams very difficult

3. Single-dimensional analytics not enough
   - Must consider network and content both
   - Aggregation of multiple posts/documents for robust analysis
   - Use and integrate different analytic models

4. Human needs to be in the loop
   - Black box systems can suggest but not decide
   - Human intuition needed to guide analytics
   - Interfaces needed to help guide process
My focus is on combining network and content

- Much work in networks and content separately
- Combining them rarely seen in social media mining
  - Machine learning and data mining has statistical relational models (graphical models), but they do not scale well
- Big data problem
  - Dimensions: realtime, volume, heterogeneity
  - The world changes rapidly: what is going on right now
  - → slow models may not be the best
- Answer: start small, then build up
  - Focus on real data from the beginning is key
    - Handle data access, noise, missing data, etc. from the get-go
Agenda

• User profiling from posts
  • Is it possible?
  • What can they be used for
  • When and why to use different representations

• Can we characterize different types of posts?
  • Social dialogues
  • Topics across different types of posts
Recent work

- **Blogosphere**
  - Semantically tagging of links [Macskassy 2010, 2011]
  - Gained new insights into evolution of topics not possible before
  - **Demographics** [Michelson and Macskassy, 2011]
    - Could extract demographic information from blog-posts
    - Demographic clusters and network clusters very different!
    - Could identify demographic characteristics of web-sites based on which bloggers linked to them (in)directly

- **Twitter**
  - User interests based on tweets [Michelson and Macskassy, 2010]
  - Model retweeting behavior [Michelson and Macskassy, 2011]
  - Representation matters [Macskassy, 2012]
  - What do people talk about?
Twitter

Johncrossmirror

#Arsenal superb, Cesc magnificent and a brilliant victory. Dunno where the balance is of Arsenal being brilliant and Braga being awful...

Off to #Arsenal shortly for Champs League opener with Sporting Braga. Prediction free zone tonight!

@CallCollymore Sure you already have it, but new Arcade Fire is immense. Have listened to nothing else since I bought it

#Arsenal striker Bendtner insists injury hell coming to an end and his views on Braga: http://tinyurl.com/33feq2f
How to represent tweets?

- **Problem with tweets**
  - They are short (max 140 chars)
  - They contain bad grammar, misspellings, etc.
  - Little context, potentially many words with little overlap
  - Topics are at term level (e.g., Arsenal) not category level (e.g., Football in England)
    - Harder for high level search and clustering → no well-defined topics

- **Idea: Leverage entities mentioned**
  - From a small sampling of user accounts:
    - ~20% of analyzed Twitter accounts mentioned 50 brands
    - 85% of trending topics are news (likely contain entities)
How to represent user profiles?

Hypothesis

• Analyzing all of users’ Tweets can yield their interests
  – Specifically, focus on Named-Entities and the topics they “represent”
  – These topics = the users’ topics of interest

#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc

User likes topics related to English soccer, international soccer, ...
Discover Entities in Tweets

Named Entity Extraction in Tweets

- All caps, all lowercase
- Not often parseable (e.g., POS-tags)
- Look for capitalized (non-stop) words

#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc
Disambiguate: Issue

Goal:

- Look up in Wikipedia → get categories → “topics of interest” for user

#Arsenal  winger  Walcott:  Becks  is my England inspiration: http://tinyurl.com/37zyjsc

~50 possibilities
Arsenal (Kremlin)
Foochow Arsenal
Arsenal Street
Arsenal F.C. (England)
Arsenal F.C. (Argentina)
Arsenal (Basketball)
Arsenal (Comic)
Arsenal (film)
Arsenal (automobile)
...

~15 possibilities
Walcott, Lincolnshire (England)
Walcott, Iowa (U.S.)
Clyde Walcott
Derek Walcott
Theo Walcott
Mary Walcott
...

3 possibilities
Beck’s Brewery
David Beckham
Beckett Scott
Disambiguate: Context

Leverage “context” of Tweet to aid disambiguation

**Entity:** Arsenal

**Context:** \{winger, Walcott, Becks, England, inspiration, ...\}

Language model:
Maximize similarity

$$\underset{e_i \in E}{\text{arg max}} \left( C_T \cap C_{e_i} \right)$$

**Arsenal = Arsenal F.C.**
Folksonomy: “Theo Walcott”

Categories: 1989 births | Living people | People from Stanmore | English footballers

Categories: English sportspeople | Association football players by nationality | Football in England | British footballers

Categories: Association football players | Sportspeople by sport and nationality | Association football by country
**Topic Profiles**

**After Step 1: Forest of category trees per Tweet**

Weight of category $c$ for user $u$:

$$w_c(u) = \sum_{t \in T_u} b^{-d_{c,t}}$$

Represent as a vector of (category-id:weight) elements:

$$\{(\text{cat}3,0.5), (\text{cat}62,5.4), (\text{cat}298,0.2), \ldots, (\text{cat}43876,1.7)\}$$
Getting a user profile from generated content

Input Tweets for User

#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/3zyjyj

Oh, and Senderos’s Dad is lovely fella too. Philippe cried in d/r after #Arsenal CL defeat at Liverpool. That shows he cares - wish more did

Step 1: Discover Categories
- Discover Entities in Tweet
- Disambiguate Entities
- Retrieve Folksonomy Sub-Tree

Step 2: Discover Profile
- Generate Topic Profile from Sub-Trees

Topic Profile:
“English Football”
“World Cup”
Proof-of-concept tiny study: Data

2 “known” Twitter users [know topics of interest]
  • johncrossmirror → soccer writer for UK newspaper
  • gizmodo → technology blog (loves Apple!)
  • Validity for users with specific topics, lots of NE

2 random Twitter users
  • Read Tweets by hand to see topics of interest.
  • Validity for “random” users

<table>
<thead>
<tr>
<th>User</th>
<th># Tweets collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johncrossmirror</td>
<td>280</td>
</tr>
<tr>
<td>Gizmodo</td>
<td>599</td>
</tr>
<tr>
<td>Anonymous1</td>
<td>340</td>
</tr>
<tr>
<td>Anonymous2</td>
<td>180</td>
</tr>
</tbody>
</table>
Results: Precision@K

Read posts and examine generated topics

- “Relevant” or “Not Relevant”

<table>
<thead>
<tr>
<th>Size of Top K</th>
<th>Avg. Precision (± Std. Dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.95 ± 0.10</td>
</tr>
<tr>
<td>10</td>
<td>0.90 ± 0.08</td>
</tr>
<tr>
<td>25</td>
<td>0.85 ± 0.08</td>
</tr>
<tr>
<td>John Cross</td>
<td>Gizmodo</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>UK soccer writer</td>
<td>Tech blog (+Apple)</td>
</tr>
<tr>
<td>ASSOCIATION FOOTBALL PLAYERS</td>
<td>COMMUNICATION</td>
</tr>
<tr>
<td>2010 FIFA WORLD CUP PLAYERS</td>
<td>APPLE INC.</td>
</tr>
<tr>
<td>SPORT IN ENGLAND</td>
<td>EMBEDDED SYSTEMS</td>
</tr>
<tr>
<td>FOOTBALL IN ENGLAND</td>
<td>COMPANIES ESTABLISHED IN 1976</td>
</tr>
<tr>
<td>2006 FIFA WORLD CUP PLAYERS</td>
<td>COMPANIES BASED IN CUPERTINO, CALIFORNIA</td>
</tr>
<tr>
<td>SPORTS TEAMS BY COUNTRY</td>
<td>TECHNOLOGY</td>
</tr>
<tr>
<td>ASSOCIATION FOOTBALL IN EUROPE</td>
<td>TELECOMMUNICATIONS</td>
</tr>
<tr>
<td>ORGANISATIONS BASED IN ENGLAND</td>
<td>MEDIA TECHNOLOGY</td>
</tr>
<tr>
<td>ASSOCIATION FOOTBALL</td>
<td>COMPUTING</td>
</tr>
<tr>
<td>SPORT IN ENGLAND BY SPORT</td>
<td>ELECTRONIC HARDWARE</td>
</tr>
</tbody>
</table>
Discussion

Other topic models (why not LDA?)

• Data is sparse (not many Tweets, they are short)
• Topics are
  • Term level (e.g., Arsenal)
  • Not category level (e.g., Football in England)
  • Harder for high level search and clustering → not well defined topics

Leveraging entities

• ~20% of analyzed Twitter accounts mentioned 50 brands
• 85% of trending topics are news (likely contain entities)
Discussion

Hashtags

• User-given token for search and categorization

<table>
<thead>
<tr>
<th>Username</th>
<th>Hashtags (ordered)</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Cross</td>
<td>Arsenal, England, wc2010, Spurs, mufc</td>
</tr>
<tr>
<td>Gizmodo</td>
<td>iPad, Apple, memoryforever, ipadapps, photography</td>
</tr>
<tr>
<td>Anonymous1</td>
<td>USC, dadt, glee, omgfacts, spoileralert</td>
</tr>
<tr>
<td>Anonymous2</td>
<td>Cubs, Nowplaying, Blackhawks, Chicago, MLB</td>
</tr>
</tbody>
</table>

• At term level and some are ill-defined
  • Overly specific, difficult to analyze, short life-span, ...
  → Hard to use as topics of interest
Agenda

• User profiling from posts
  • Is it possible? (demo at end)
  • What can they be used for
  • When and why to use different representations

• Can we characterize different types of posts?
  • Social dialogues
  • Topics across different types of posts
Question: What makes People Retweet?

• Twitter is an interesting social media broadcast platform
  • Twitterers can broadcast small messages
• Twitter has introduced syntactic constructs to help navigate tweets
  • Retweeting (re-broadcast a message):
    • [user_K] I just saw Kevin Bacon at the Kids’R’Us in La Brea!
    • [user_X] RT @user_K I just saw Kevin Bacon at the Kids’R’Us in La Brea!
  • “Messaging”
    • [user_K] @user_X Hey, stop retweeting me!
  • “Topics”
    • [user_K] #KevinBacon He just let Kids’R’Us with a huge doll house!
• The question we ask here is why do users retweet?
  • Is it based on things they like?
  • Is it based on who posts?
Hypothesis

• Knowing more about a user’s interests and past behavior can help predict his or her future retweets

• We will develop and test three hypotheses
  • Homophily: Users are more likely to retweet information coming from people who are like them
    → Similarity of two profiles
  • Topic: Users are more likely to retweet information they find interesting
    → Similarity of user profile to tweet
  • Network: Users are more likely to retweet information from people they were in recent communications with

• Null hypothesis: Random model
Retweet models

• Null model (General “recency” model)

\[ P_{gm}(x) = c \cdot \text{time}(x)^{-\alpha} \]

• Networking

\[ P_{\text{network}}(x) = P_{gm} \cdot \left[ \alpha \cdot P(x | \text{recent}(x)) + (1 - \alpha) \cdot P(x | \neg \text{recent}(x)) \right] \]

• Topic

\[ P_{\text{topic}}(x) = P_{gm} \cdot P_{ts}(x | \text{sim}_{\text{topic}}(x,u)) \]

• Homophily

\[ P_{\text{homophily}}(x) = P_{gm} \cdot P_{ps}(x | \text{sim}_{\text{homophily}}(x,u)) \]
Computing similarity

- We use a standard cosine distance metric to compute similarities

\[ sim(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|} \]
Retweet Study: Data

Collected 4 weeks of tweets from ~30K Twitterers Using geographic-based snowball sampling

• Selected seed ~200 Twitterers in Pakistan and Israel
  • Extracted users mentioned retweet or messages
  • Added users who were (self-reportedly) in Pakistan or Israel

• Increased to ~30K Twitterers in a matter of a week

Using tweets from users from 9/20/10 to 10/20/10

• 768K tweets
Retweet Study: Data (cont’d)

Consider only users who had 3+ tweets and 3+ retweets

- 482K tweets (43% have concepts; 84% have words)
- 103K retweets (70% have concepts; 94% have words)
- 16K retweets of 1800 users where both original tweeter and retweeter had 3+ tweets and 3+ retweets
Fitting the models / Evaluation

- We have four models
- We fit the parameters to the data and evaluate which model best fits the data
How much do users retweet? (log-log-scale)

How much do people retweet?
→ 30% of tweets are retweets (log-log scale)

What is the recency of retweets?
Tend to retweet recent tweets
Powerlaw: $\alpha=1.15$
(log-log scale)

Null Model
Retweet behavior: What does the data tell us?

Homophily: More likely to retweet someone similar? (y log-scale)

\[ P_{psim}(x \mid \text{sim}_p(x,u)) \]

Topic: More likely to retweet something interesting? (y log-scale)

\[ P_{t\text{sim}}(x \mid \text{sim}_T(x,u)) \]

Network: 38% of retweets to someone recently (<1 day) retweeted
60% of retweets are to someone previously retweeted (x log-scale)

\[ \Rightarrow \text{Use logodds ratio at given similarity percentile} \]
Retweet models

- Null model (General “recency” model)
  \[ P_{gm}(x) = 0.2 \cdot \text{time}(x)^{-1.15} \]

- Networking (\(\alpha = 0.38\))
  \[ P_{network}(x) = P_{gm} \cdot \left[ \alpha \cdot P(x \mid \text{recent}(x)) + (1 - \alpha) \cdot P(x \mid \neg\text{recent}(x)) \right] \]

- Topic
  \[ P_{topic}(x) = P_{gm} \cdot P_{ts}(x \mid \text{sim}_{topic}(x,u)) \]

- Homophily
  \[ P_{homophily}(x) = P_{gm} \cdot P_{ps}(x \mid \text{sim}_{homophily}(x,u)) \]
Evaluation

• Questions we ask:
  • Globally: Which models fits best?
    • Might be dominated by prolific users (10% generating 90% of traffic)
  • By user: Which model fits best?
    • Might not get diversity of users
  • By tweet: How often is each model used per user?

• Second, does the representation have an effect?
  • Concepts vs. text vs. hybrid

• Finally: what do the networks look like? Can we not just look at the follower-graph to mine diffusion?
Evaluation using concept representation

_Globally_
Homophily (45% of all retweets)

_By User_
Ratio of users explained by each model

<table>
<thead>
<tr>
<th></th>
<th>Null Model</th>
<th>Network</th>
<th>Homophily</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>12%</td>
<td>14%</td>
<td>67%</td>
<td>31%</td>
</tr>
</tbody>
</table>

On average (by tweet), users used the following models

<table>
<thead>
<tr>
<th></th>
<th>Null Model</th>
<th>Network</th>
<th>Homophily</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>11%</td>
<td>26%</td>
<td>37%</td>
<td>26%</td>
</tr>
</tbody>
</table>
Evaluation using concept representation

**By Tweet**

On average, users used the following models

- User’s behaviors actually best explained by a mixture of models
- How many many models needed to explain all of a user’s behavior?

<table>
<thead>
<tr>
<th></th>
<th>Null Model</th>
<th>Network</th>
<th>Homophily</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11%</td>
<td>26%</td>
<td>37%</td>
<td>26%</td>
</tr>
</tbody>
</table>

![Bar chart showing how many models are used](chart.png)
Agenda

• User profiling from posts
  • Is it possible? (demo at end)
  • What can they be used for
  • When and why to use different representations

• Can we characterize different types of posts?
  • Social dialogues
  • Topics across different types of posts
What about using a text-based profile?

• We use a standard **tf.idf** model, where we consider all tweets (not retweets) from a person as one long document
• We do **stemming** and **stop-listing** to reduce dimensionality
• We then represent a profile of a user as a standard bag-of-words tf.idf vector of (wordID, wt) elements:

Represent as a vector of (word-id:weight) elements:
{ (wing,0.5), (soccer,5.4), (beck,0.2), ... , (uefa,1.7)}

Weights are computed using the Okapi BM25 ranking:

\[
\begin{align*}
    w_w(u) &= IDF(w) \cdot \frac{f(w,u) \cdot (k_1 + 1)}{f(w,u) + k_1 \cdot \left(1 - b + b \cdot \frac{|D_u|}{\text{avgD}}\right)} \\
    IDF(w) &= \log \frac{N_u - n(w) + 0.5}{n(w) + 0.5}
\end{align*}
\]

We prune rare words (<5 users use them)}
Computing similarity

• We use a standard cosine distance metric to compute similarities

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|}$$

• This works equally well when comparing text vectors or category vectors

• Can also get a hybrid score by taking the average of the text and category similarities

➢ So we can now test our hypotheses across the three models as well as across 3 representations
Evaluation: Which representation to use?

• We have three types of representations
  • Concepts
  • Text
  • Hybrid (averaging the above two)

• Does representation have an effect on fit?
  • Some tweets may not have concepts
    • Therefore words might be better
  • Some tweets on same concept may not share words
    • Therefore using concepts might be better
  • A hybrid might be the best of both worlds?
How does text- and category-profiles correlate?

- How do similarity scores correlate between the word and concept representations?
Retweet Study: Data

• Collected 4 weeks of tweets from ~30K Twitterers
• Using geographic-based snowball sampling
• Using tweets from users from 9/20/10 to 10/20/10
  • 768K tweets
• Include users who had 3+ tweets and 3+ retweets
  • 482K tweets (43% have concepts; 84% have words)
  • 103K retweets (70% have concepts; 94% have words)
  • 16K retweets of 1800 users where both original tweeter and retweeter had 3+ tweets and 3+ retweets
**Global** evaluation across different representations

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>1446</td>
<td>9%</td>
</tr>
<tr>
<td>Network</td>
<td>4918</td>
<td>32%</td>
</tr>
<tr>
<td>Homophily</td>
<td>7037</td>
<td>45%</td>
</tr>
<tr>
<td>Topic</td>
<td>3183</td>
<td>20%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>946</td>
<td>6%</td>
</tr>
<tr>
<td>Network</td>
<td>4976</td>
<td>32%</td>
</tr>
<tr>
<td>Homophily</td>
<td>6486</td>
<td>42%</td>
</tr>
<tr>
<td>Topic</td>
<td>3834</td>
<td>25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>1005</td>
<td>6%</td>
</tr>
<tr>
<td>Network</td>
<td>5229</td>
<td>34%</td>
</tr>
<tr>
<td>Homophily</td>
<td>6591</td>
<td>42%</td>
</tr>
<tr>
<td>Topic</td>
<td>3390</td>
<td>22%</td>
</tr>
</tbody>
</table>
By user evaluation across representations

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>203</td>
<td>12%</td>
</tr>
<tr>
<td>Network</td>
<td>234</td>
<td>14%</td>
</tr>
<tr>
<td>Homophily</td>
<td>1116</td>
<td>67%</td>
</tr>
<tr>
<td>Topic</td>
<td>518</td>
<td>31%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>145</td>
<td>9%</td>
</tr>
<tr>
<td>Network</td>
<td>222</td>
<td>13%</td>
</tr>
<tr>
<td>Homophily</td>
<td>1076</td>
<td>65%</td>
</tr>
<tr>
<td>Topic</td>
<td>597</td>
<td>36%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Wins</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>145</td>
<td>9%</td>
</tr>
<tr>
<td>Network</td>
<td>232</td>
<td>14%</td>
</tr>
<tr>
<td>Homophily</td>
<td>1143</td>
<td>69%</td>
</tr>
<tr>
<td>Topic</td>
<td>529</td>
<td>32%</td>
</tr>
</tbody>
</table>

- **Text** works better for **Topic** model because individual tweets have less data
- **Concepts** work better for **Homophily**-model because high-level concepts better captures a user’s interest
- **Hybrid** representation had *best overall* performance for Homophily model
Agenda

• User profiling from posts
  • Is it possible? (demo at end)
  • What can they be used for
  • When and why to use different representations

• Can we characterize different types of posts?
  • Social dialogues
  • Topics across different types of posts
Extracting Dialogues

<table>
<thead>
<tr>
<th>Time</th>
<th>User</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>23:56</td>
<td>User1</td>
<td>@User2 why don’t you get a car my friends</td>
</tr>
<tr>
<td>00:00</td>
<td>User2</td>
<td>@User1 cause my cars transmission blew before i</td>
</tr>
<tr>
<td>00:01</td>
<td>User1</td>
<td>@User2 ohh and when you come back we must go</td>
</tr>
<tr>
<td>00:03</td>
<td>User2</td>
<td>@User1 DEFINITELY im going there and in an o</td>
</tr>
<tr>
<td>00:09</td>
<td>User1</td>
<td>@User2 do you not have a in n out to??</td>
</tr>
<tr>
<td>00:16</td>
<td>User2</td>
<td>@User1 no we do but its hella far :( i come back i</td>
</tr>
<tr>
<td>00:19</td>
<td>User1</td>
<td>@User2 my birthday!! I’ll drive??</td>
</tr>
<tr>
<td>00:23</td>
<td>User2</td>
<td>@User1 im sooo down..my parents wanna get me</td>
</tr>
</tbody>
</table>

Table 3: Example two-way dialogue

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>User</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>18:55</td>
<td>User1</td>
<td>@User7 MQM is THE MAFIA! The organiz</td>
</tr>
<tr>
<td>(A)</td>
<td>18:57</td>
<td>User1</td>
<td>@User2 bro, please stop misstating me. I love my children</td>
</tr>
<tr>
<td>(A)</td>
<td>19:00</td>
<td>User2</td>
<td>@User1 Mafia of MQM makes 70% of Karachi</td>
</tr>
<tr>
<td>(B)</td>
<td>19:01</td>
<td>User3</td>
<td>@User4 @User5 @User2 v hope 4 a political</td>
</tr>
<tr>
<td>(A)</td>
<td>19:04</td>
<td>User1</td>
<td>@User2 70% of Karachi is MQM? Really??</td>
</tr>
<tr>
<td>(C)</td>
<td>19:05</td>
<td>User2</td>
<td>@User6 @User1 @User5 @User3 MQM</td>
</tr>
<tr>
<td>(C)</td>
<td>19:06</td>
<td>User2</td>
<td>@User1 ok then app battaa do... Laikin baat</td>
</tr>
<tr>
<td>(C)</td>
<td>19:09</td>
<td>User4</td>
<td>@User3 @User5 @User2 :) No Maseeha or</td>
</tr>
<tr>
<td>(C)</td>
<td>19:10</td>
<td>User2</td>
<td>@User5 @User6 @User1 @User3 wrong,</td>
</tr>
<tr>
<td>(C)</td>
<td>19:11</td>
<td>User2</td>
<td>@User4 @User3 @User5 no doubt about t</td>
</tr>
<tr>
<td>(C)</td>
<td>19:13</td>
<td>User1</td>
<td>@User2 :) Your 'facts’ tell me this discussion.</td>
</tr>
</tbody>
</table>
How do people spend their time?

Activity profiles of users: what fraction of users spend 0% through 100% of their time posting each type of tweet?

<table>
<thead>
<tr>
<th>Tweet Category</th>
<th>Number Tweets</th>
<th>Overall Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue</td>
<td>66,812</td>
<td>0.13</td>
</tr>
<tr>
<td>Retweet</td>
<td>93,319</td>
<td>0.19</td>
</tr>
<tr>
<td>Mention</td>
<td>154,177</td>
<td>0.31</td>
</tr>
<tr>
<td>Tweet</td>
<td>183,748</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>498,056</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>Number</th>
<th>Ratio</th>
<th>Avg. Num. Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>18,619</td>
<td>92.37%</td>
<td>4.9</td>
</tr>
<tr>
<td>3</td>
<td>1,232</td>
<td>6.11%</td>
<td>8.5</td>
</tr>
<tr>
<td>4</td>
<td>181</td>
<td>0.90%</td>
<td>12.7</td>
</tr>
<tr>
<td>5</td>
<td>83</td>
<td>0.41%</td>
<td>19.4</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>0.13%</td>
<td>36.5</td>
</tr>
<tr>
<td>&gt; 6</td>
<td>13</td>
<td>0.07%</td>
<td>&gt; 60</td>
</tr>
</tbody>
</table>
Tweet characteristics by class
How is user attention split across friends?

Entropy for user $u$:

Over dialogues:
$$- \sum_{n \in N^u} d_{u,n} \log(d_{u,n}) + \sum_{n \in N^u} (1 - d_{u,n}) \log(1 - d_{u,n})$$

$d_{u,n} = \text{fraction of } u\'s\ \text{dialogues which include } n$

Over tweets:
$$- \sum_{n \in N^u} r_{u,n} \log(r_{u,n}) + \sum_{n \in N^u} (1 - r_{u,n}) \log(1 - r_{u,n})$$

$r_{u,n} = \text{fraction of } u\'s\ \text{tweets which include } n$

Takeaway:
Dialogues cover many people, but overall users interact directly with only a few
How many people *really* participate?

Entropy for dialogue $D$:

$$- \sum_{u \in D} r_{u,D} \log(r_{u,D}) + \sum_{u \in D} (1 - r_{u,D}) \log(1 - r_{u,D})$$

$r_{u,D} =$ fraction of $D$’s tweets which originated with $u$

**Takeaway:**

The more users participating in a dialogue, the more likely that a few people dominate the discussion.
Agenda

• **User profiling from posts**
  • Is it possible? (demo at end)
  • What can they be used for
  • When and why to use different representations

• **Can we characterize different types of posts?**
  • Social dialogues
  • Topics across different types of posts
What are people talking about?

Different types of posting behaviors

• Retweeting: *information diffusion*  
  [Macskassy and Michelson 2011; Macskassy 2012]

• Social dialogues/chat: *social networking*  
  [Macskassy 2012]

• General: *Broadcast* tweets to all followers

*How do they differ in terms of topics?*
Topic Study: Data

• Using tweets from users from 9/20/10 to 10/20/10
  • 768K tweets

• Categorize tweets into dialogues, retweets and other
  • Dialogues: 108K tweets
  • Retweets: 116K tweets
  • Other (general): 429K tweets
**Topic Study: Methodology**

- Used LDA topic-model
  - Identified 150 topics per tweet category
  - 450 total topics
- Manually labeled 450 topics
  - 80 emergent categories
- Counted how often each category was present in each tweet category
# Topics for each kind of behavior

## Dialogues (108K)

<table>
<thead>
<tr>
<th>Categories (37)</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-talk</td>
<td>45.8%</td>
</tr>
<tr>
<td>Daily life</td>
<td>16.4%</td>
</tr>
<tr>
<td>Twitter/FB/...</td>
<td>12.3%</td>
</tr>
<tr>
<td>Justin Bieber</td>
<td>10.3%</td>
</tr>
<tr>
<td>School</td>
<td>6.9%</td>
</tr>
<tr>
<td>Music</td>
<td>5.3%</td>
</tr>
<tr>
<td>Complaining</td>
<td>4.8%</td>
</tr>
<tr>
<td>Sports</td>
<td>4.1%</td>
</tr>
<tr>
<td>Work</td>
<td>3.7%</td>
</tr>
<tr>
<td>TV</td>
<td>3.8%</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

## General tweets (429K)

<table>
<thead>
<tr>
<th>Categories (43)</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-talk</td>
<td>17.5%</td>
</tr>
<tr>
<td>Breaking news (event)</td>
<td>8.1%</td>
</tr>
<tr>
<td>Complaining</td>
<td>6.8%</td>
</tr>
<tr>
<td>Technology</td>
<td>6.7%</td>
</tr>
<tr>
<td>News</td>
<td>6.2%</td>
</tr>
<tr>
<td>Politics</td>
<td>5.4%</td>
</tr>
<tr>
<td>Twitter/FB/...</td>
<td>5.3%</td>
</tr>
<tr>
<td>Sports</td>
<td>4.7%</td>
</tr>
<tr>
<td>Daily life</td>
<td>4.1%</td>
</tr>
<tr>
<td>Religion</td>
<td>4.0%</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

## Retweets (116K)

<table>
<thead>
<tr>
<th>Categories (45)</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>14.4%</td>
</tr>
<tr>
<td>Small-talk</td>
<td>9.3%</td>
</tr>
<tr>
<td>Breaking news (event)</td>
<td>8.3%</td>
</tr>
<tr>
<td>Politics</td>
<td>7.8%</td>
</tr>
<tr>
<td>Technology</td>
<td>6.7%</td>
</tr>
<tr>
<td>Justin Bieber</td>
<td>5.9%</td>
</tr>
<tr>
<td>Religion</td>
<td>4.9%</td>
</tr>
<tr>
<td>Pakistan</td>
<td>4.5%</td>
</tr>
<tr>
<td>Music</td>
<td>4.4%</td>
</tr>
<tr>
<td>Sports</td>
<td>4.2%</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

### Color Legend:

- **Social / Personal topic: (life)** “I didn’t have time to buy groceries today.”
- **Public topic: (breaking news):** “The flood has now destroyed 100 homes”
- **Smalltalk:** “Hey, it’s been a while. How are you doing?”
Agenda

- User profiling from posts
  - Is it possible? (demo at end)
  - What can they be used for
  - When and why to use different representations

- Can we characterize different types of posts?
  - Social dialogues
  - Topics across different types of posts
Summary

• Social media analytics is a rich domain for many different mining and learning technologies
  • Technologies developed here are broadly applicable
  • While joint models likely better, improvements in core technologies crucial to move field forward

• Today I focused primarily on text mining and its uses
  • User profiling and underlying representations
    • Text is good when analyzing single tweets
    • Higher-level mapping more salient for aggregate analysis
  • Explaining retweeting behaviors
  • Homophily rules the day
  • Clear difference in topics for different categories of posts

• Still need a lot of work!
Demo

https://github.com/InformationIntegrationGroup/EntityExplorer
Social Media Analytics: Text Mining for User Profiles and More

https://github.com/InformationIntegrationGroup/EntityExplorer

AAAI has a nice Symposium on Data Mining and Text Analytics. Experts in AI and NLP are here.

Identify Entities

7 entities found

<table>
<thead>
<tr>
<th>Wikipedia Entity Name</th>
<th>Use in Graph</th>
<th>Wikipedia Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy text analytics</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>Data mining</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>AAAI Conference on Artificial Intelligence</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>Studies in NLP</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>Experts-Exchange</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>AI winter</td>
<td></td>
<td>View Categories</td>
</tr>
<tr>
<td>Symposium</td>
<td></td>
<td>View Categories</td>
</tr>
</tbody>
</table>
Entity's Wikipedia Category visualization

Data mining

- Algorithms and data structures
- Computer programming
- Algorithms
- Mathematical concepts
- Mathematical problem solving

- Information science

- Data analysis

- Statistics
- Mathematical sciences
- Mathematical and quantitative methods (economics)

- Analysis

- Formal sciences
- Data management

https://github.com/InformationIntegrationGroup/EntityExplorer
Thank you