Research Overview

Arjun Mukherjee
University of Houston
September 30, 2014
Goal

(1) Mine and analyze social media + user’s interaction with the Web

(2) Characterize user behaviors, language and sentiments in social media to predict socio-psychological indices:

- Gender, age, mood, agreement/disagreement etc.
- Socio-economic indices (e.g., stock markets)
- Deception and Fraud (e.g., fake review, identities, sockpuppets)

Using language and user behaviors in social media.
Research Areas

- Web/Data Mining
- Machine Learning
- Natural Language Processing

Sentiment Analysis
Computational social science
Filter Opinion Fraud/Spam
Outline

- Introduction – Motivation
- Problem#1: Modeling debate conversations [KDD’12; ACL’13]
- Problem#2: Social Sentiment ⇒ Stock Prediction [ACL’13]
- Problem#3: Deceptive opinion fraud/spam [KDD’13, WWW’12]
Modeling social debates

Input: 1.2 M non-stop words, 100,000+ debate posts

Output: Latent topics, A/D sentiment expressions

Topics
- Global warming
- Earth
- CO2
- Vegetarianism
- Animals
- Meat
- Beef
- Slaughter
- Kill
- Life
- Diet
- Nuclear waste
- Sea level
- Climate change
- Arctic ice
- Meat industry
- Vegan
- Vegetables
- Meat consumption
- Tax
- Government
- IRS/Taxes
- Money
- Pay
- Federal
- State taxes
- Services
- Social security
- Income tax
- Budget
- Vote
- President
- Democrats
- Politics
- Electoral
- Obama
- Government
- Policy
- Elections
- Senate
- Libertarian
- Left wing

A/D Sentiments
- Agreement
  - I agree
  - Indeed correct
  - I accept
  - Claim
  - I concede
  - Is valid
  - I do support
  - Rightly said
  - Completely agree
  - Very well put
  - Point taken

- Disagreement
  - I disagree
  - Reject
  - I would disagree
  - I don’t buy your
  - Really
  - I really doubt
  - Can you prove
  - Argument fails
  - You fail to
  - Contradict yourself
  - Don’t understand

Arjun Mukherjee
Why do we care?

• Q: How do people feel about taxes, terrorism?

• Useful for policy, bills, campaigns

• Data: Online debate/discussions (Volconvo.com)

30-Sep-14

Arjun Mukherjee
Outline

- Introduction – Motivation
- Problem#1: Modeling debate conversations [KDD’12; ACL’13]
- Problem#2: Social Sentiment ⇒ Stock Prediction [ACL’13]
- Problem#3: Deceptive opinion fraud/spam [KDD’13, WWW’12]
Decision Making using Social Sentiment

• Stock tweets
  “$AAPL is loosing customers. everybody is buying android phones! $GOOG.”
  “I think $AAPL is signaling a huge drop that will come in the mkt…”

• Dedicated message boards:
  Stocktwits, marketwatch, Yahoo! Finance

• Social sentiment ↔ Stock Market

• Q: How to use social sentiment to make decisions?
An Interesting Pattern

Daily Prediction: Dow Jones Industrial Average (DIA) v.s. Twitter Sentiment
Source: Yahoo! Finance & Twitter.com

- Pos/Neg sentiment ratio’s trend dovetails with market’s trend
- Social Mood ↔ Financial Gains
- Q: Can we do better?

30-Sep-14
Arjun Mukherjee
Introduction – Motivation

Problem#1: Modeling debate conversations [KDD’12; ACL’13]

Problem#2: Social Sentiment ⇒ Stock Prediction [ACL’13]

Problem#3: Deceptive opinion fraud/spam [KDD’13, WWW’12]
The Opinion Curse – Opinion Spam!

- Opinions ⇒ Virtual Currency

- Opinion Spam: Illegitimate activities (e.g., writing fake reviews/ratings) to deliberatively mislead consumers

- In E-commerce, filtering opinion spam is vital
Group Opinion Spam on Amazon.com

<table>
<thead>
<tr>
<th>Product</th>
<th>Date</th>
<th>Reviewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Xtract (CD-ROM)</td>
<td>December 4, 2004</td>
<td>Practically FREE music</td>
</tr>
<tr>
<td>Pond Aquarium 3D Deluxe Edition</td>
<td>December 4, 2004</td>
<td>My kids love it,</td>
</tr>
<tr>
<td>Audio Xtract (CD-ROM)</td>
<td>December 8, 2004</td>
<td>Like a tape recorder...</td>
</tr>
<tr>
<td>Audio Xtract Pro (CD-ROM)</td>
<td>December 8, 2004</td>
<td>This software really rocks.</td>
</tr>
<tr>
<td>Pond Aquarium 3D Deluxe Edition</td>
<td>December 8, 2004</td>
<td>For the price you...</td>
</tr>
<tr>
<td>Audio Xtract (CD-ROM)</td>
<td>December 4, 2004</td>
<td>Wow, internet music!...,</td>
</tr>
<tr>
<td>Audio Xtract Pro (CD-ROM)</td>
<td>December 4, 2004</td>
<td>Best music just got....</td>
</tr>
<tr>
<td>Pond Aquarium 3D Deluxe Edition</td>
<td>December 4, 2004</td>
<td>Cool, looks great...,</td>
</tr>
</tbody>
</table>

1. All reviewed same 3 products giving all 5-stars + 100% helpfulness votes!
2. All reviews posted in a time window of 4 days
3. Each only reviewed those 3 products,
4. **Unlikely to be coincidental- Something seems fishy!**

Arjun Mukherjee
Research Areas I work on…

- Join me as an RA!
- Email for more info: arjun@cs.uh.edu