The Journey of My Research

IST Brown Bag Seminar

John Yen
March 25, 2015
Collaborators

**Cancer Informatics Initiative**
Prasenjit Mitra
Vasant Honavar
David Reitter
Michelle Newman
Heath Mackley
Ngot Bui
Yafei Wang
Prakhar Biyani
Mo Yu
Cornelia Caragea (U. N. Texas)
Kang Zhao (U. Iowa)
Lior Rokach (BGU)
Nir Ofek (BGU)
Baojun Qiu (Straits)
Siddhartha Banerjee
Yang Xu

**American Cancer Society**
Kenneth Portier
Greta Greer

**Cyber Security**
Peng Liu
Cheng Zhong
Gaoyao Xiao
Ping Li
Nanying Zhang
Frank Ritter
Gerry Santoro
Army Research Lab
Robert Erbacher
Steve Hutchinson
Christopher Garneau
Hasan Cam
William Glodek
My childhood passion is to become an Electrical Engineer

- 1960’s-1970’s, Hsinchu, Taiwan
The Beginning of my Journey toward a “Scientist”

• 2010, American Cancer Society (ACS), Atlanta, GA
• Collaborated on a proposal regarding network analysis of the Cancer Survivors Network (CSN) operated by ACS
• Visited ACS with Prasenjit
• Met Greta Greer and Kenneth Portier regarding possible collaborations

• Top questions ACS is interested in
  – How to better understand the benefits of CSN members?
  – How to identify influential users?
These two questions turned out to be important for online health community (OHC)

Adult Internet Users in the U.S. (Pew)

80% use Internet for health-related purposes

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

Cancer Survivors Network
Previous Research on Benefits of OHC

• Through surveys revealed that members
  – Received more social support (Dunkel-Schetter 1984, Rodgers and Chen, 2005)
  – Reduced the level of stress, depression, and psychological trauma (Beaudoin and Tao 2008, Winzelberg et al, 2003)
  – Became more optimistic about disease (Rodgers and Chen, 2005)

• Q1-1: Can we use computational text analysis to capture the changes of their mood over time?
Computational Analysis of Emotion

• Bag-of-Word approach
  – Occurrence of words in a word list associated with positive/negative emotion
  – LIWC (Linguistic Inquiry and Word Count) (J. W. Pennebaker)
  – Occurrence of words in a word list associated with strength of positive/negative emotion

• Part-of-Speech (POS)
  – Syntactic patterns of sentences
  – Useful also to extract what the sentiment is about

• Unsupervised Learning (Liu, 2010)

• Supervised Learning facilitates “context” to be incorporated into these models.

Sentiment Classifier Using Supervised Machine Learning

- **Model training**
  - Manually labeled 298 posts (-5 to 5)
  - 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 untrained labeled posts (cross validation)

- **Model prediction**
  - Apply the model to label the sentiment of unlabeled posts of the entire community
Features

### Features from modified sentiment word-lists

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

### Novel features introduced through insights
Which sentiment model to choose?

- 10-fold cross-validation
- AdaBoost-based sentiment model provided the best result
What are the key features used in the model?

<table>
<thead>
<tr>
<th>LOO Feature</th>
<th>ROC Area</th>
<th>Decrease from the full model</th>
</tr>
</thead>
<tbody>
<tr>
<td>PosStrength</td>
<td>0.774</td>
<td>0.054</td>
</tr>
<tr>
<td>PosVsNeg</td>
<td>0.804</td>
<td>0.024</td>
</tr>
<tr>
<td>Neg</td>
<td>0.813</td>
<td>0.015</td>
</tr>
<tr>
<td>Slang</td>
<td>0.813</td>
<td>0.015</td>
</tr>
<tr>
<td>NegStrength</td>
<td>0.813</td>
<td>0.015</td>
</tr>
<tr>
<td>PostLength</td>
<td>0.819</td>
<td>0.009</td>
</tr>
<tr>
<td>Name</td>
<td>0.82</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Q1: How to quantify the benefits of CSN members?

Q1-2: How to quantify the sentiment changes of those who started a thread?

A Discussion Thread

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

$$\Delta_{Pr} = \sum_{i=1}^{n_0} Pr(P_i^0)/n_0 - Pr(P^0),$$
75% of thread originators with negative initial posts became more positive later on the thread.

Q1-3: How do those sentiment changes of thread initiator progress over time on a thread?

- The Hypothesis: The sentiment changes gradually (i.e., the more replies they received, the more likely they become positive).
The “Aha” moment of a Scientist!

- 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.
Q1: How to better understand the benefits of CSN users?

• Supervised learning-based sentiment classification of threaded discussions over 10 years indicated that
  – A1-1: 75% of those who started a thread with negative sentiment became more positive later on the thread
  – A1-2: Surprise! Most of the sentiment changes of thread originators occur early on the thread.

• Can we use this finding to help answer the 2nd question (finding influential users)?
“Influential” Reply Posts are defined as

1. Reply posts that were “early” on the threads (i.e., before the first self-reply).
2. Reply posts whose sentiment is aligned with the direction of the sentiment change of the thread initiator

- Design a new “centrality” measure: The total number of influential replies posted by a user (IRR)
### IRR-based Influential User Ranking

- IRR is better than other centrality measures in predicting influential users

<table>
<thead>
<tr>
<th>Metrics</th>
<th>K=50</th>
<th>K=100</th>
<th>K=150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of threads initiated</td>
<td>0.342</td>
<td>0.439</td>
<td>0.585</td>
</tr>
<tr>
<td>Total number of posts</td>
<td>0.415</td>
<td>0.707</td>
<td>0.781</td>
</tr>
<tr>
<td>In-degree centrality</td>
<td>0.317</td>
<td>0.512</td>
<td>0.610</td>
</tr>
<tr>
<td>Out-degree centrality</td>
<td>0.390</td>
<td>0.659</td>
<td>0.780</td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.293</td>
<td>0.366</td>
<td>0.488</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.390</td>
<td>0.561</td>
<td>0.732</td>
</tr>
<tr>
<td>IRR</td>
<td>0.511</td>
<td>0.732</td>
<td>0.805</td>
</tr>
</tbody>
</table>

An Alternative Approach Using 60 Features

• Contribution features
  – The numbers of posts/threads
  – The length of posts
  – The time span of one’s activities
  – ...

• Centrality features (not IRR)
  – 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
  – ...

Compare the IRR-based approach with the alternative approach using 60+ features

- The IRR ranking (based on a single metric) outperforms the best classifier that uses 60+ metrics (ensemble classifier)

- Incorporating IRR into the classifier further improves the performance.

<table>
<thead>
<tr>
<th>Top-K recalls and precisions</th>
<th>K=50</th>
<th>K=100</th>
<th>K=150</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall (max =0.397)</td>
<td>Prec.</td>
<td>Recall (max =0.794)</td>
</tr>
<tr>
<td>The IRR Ranking</td>
<td>0.349</td>
<td>0.880</td>
<td>0.627</td>
</tr>
<tr>
<td>The ensemble classifier without IRR</td>
<td>0.278</td>
<td>0.700</td>
<td>0.532</td>
</tr>
<tr>
<td>The new ensemble classifier with IRR</td>
<td>0.373</td>
<td><strong>0.940</strong></td>
<td>0.579</td>
</tr>
</tbody>
</table>
Apple Watch

Another game changer?
Summary

• RQ1: How to better understand the benefits of CSN users?
  Supervised learning-based sentiment classification of threaded discussions over 10 years indicated that
  – 75% of those who started a thread with negative sentiment became more positive later on the thread
  – *Surprise!* Most of the sentiment changes occur early on the thread.

• RQ2: How to find influential users?
  • Early replies whose sentiment (+/-) are the same as the sentiment change (+/-) of the thread initiator is an excellent predictor of influential users.
Prima Facie Cause

• A temporal causality analysis in probabilistic computation tree logic (PCTL).
• The state C is a Prima Facie Cause of E if there is a p such that
  1. C becomes true (eventually) with a non-zero probability
  2. There exists a probability p such that reaching E after C is more likely (>p) than reaching E (through any possible paths) (<p)

How to analyze temporal causality between replies and final sentiment?

Prima Facie Causes for Final Sentiment of Thread Originators

- The positive-sentiment-replies are prima facie causes of the final positive sentiment.
- The negative-sentiment-replies are also prima facie causes of the final negative sentiment.
- For colorectal cancer community, the negative-sentiment-self-reply is also prima facie causes of the final negative sentiment.
“I am a computational social scientist?”

• Focus:
  – What causes the change of emotion state of a cancer patient in an OHC?

• Impacts:
  – Leverage this understanding to improve the supports for cancer patients

• “Reward” to self:
  – A higher degree of satisfaction
  – A broader implications (to other health conditions)
The critical point of another journey

• 2013 Summer, Adelphi, Maryland

• Goal: Conduct a study regarding cyber analysts for a MURI (PI: Peng Liu)

• Realized the threat of cyber attack is much more serious than I originally thought

• Impressed by smart people who dedicated their life and wisdom to defend the cyber space.
A “Scientist” Passion for the Human Dimension of Cyber Science

• How can we better understand the fine-grained cognitive processes of analysts who
  – need to process a large influx of network monitoring data/alerts at a rapid pace
  – need to “connect the dots” with missing links using their experience, knowledge, and skills?
An “Engineer” with a Passion for Human Dimension of Cyber Science

• Built a non-intrusive way to capture “cognitive traces”

The AOH Objects and Their Relationships in An Analyst’s Cognitive Trace
Surprise: High-performance analysts performed more filtering
How to better understand the cognitive behaviors of analysts?

• Represent each cognitive trace as a heterogeneous network
  – Node: A filtering operation on an attribute (e.g., port)
  – Link: Relationship between two filtering operations
    • Subsumption: one filtering is more specific than another ("port=4446 subsumes" “port=4446 and SourceIP= ....")
    • Complementary: one filtering complements the other
    • Equal To: identical filtering
    • Corresponding To: same condition, different data source
The Information Foraging Network of An Analyst

Nodes (Filtering)
Ordered by time around the circle.

Edges
(Relationship from a filtering to its preceding activities)
- **Orange**: Complementary
- **Red**: Equal to
- **Blue**: Subsumed by
- **Green**: Corresponding to
Both analysts have high performance score.

Their filtering networks reveal different information foraging strategies.
Figure 4. Role of working memory in explicit learning.
Am I an Engineer or a Scientist?

- Engineer Passion: Build better information value-added tools X
- Scientist Passion: Better understand the mechanism of Y

- The journey may lead to surprises (may be another “Aha” moment or ...
The Next Frontier ...
A late celebration of National Puppy Day