Project Presentations 2

CMSC 498J: Social Media Computing

Department of Computer Science
University of Maryland
Spring 2015

Cat 2: Groups G1, G2, G3, G8, G9
Presentation 2

- **G1: Going Down the Rabbit Hole**
  - Andrew, Lixian, Patrick, Andrew
- **G2: Gawker Classifier**
  - Ali, James, Jake
- **G3: Capturing a Company's Stock Performance through Twitter Sentiment**
  - Christi, Kevin, Kris, Brian
- **G8: Personal Twitter Accounts and Brand Loyalty**
  - Dylan, Gilad, Kevin, Zachery
- **G9: Twitter as a Mental Illness Diagnosis Tool**
  - Justin, Henry, Walid, Jordan
GOING DOWN THE RABBIT HOLE

Recognizing Destructive Patterns of Alcohol Abuse through Social Media

Andrew Matuza, Andrew Mawhinney, Lixian Zhou, Patrick Xiao
Who drinks?
According to the National Institute on Alcohol Abuse and Alcoholism, four out of five college students drink alcohol. Half of college students that consume alcohol admit to binge drinking. “Binge drinking is defined as a pattern of drinking that typically brings blood alcohol concentration levels to 0.08g/dL.”

Why should I care?
Engaging in high risk drinking can lead to some dire consequences. Some of these consequences include:

- Injury
- Vandalism
- Property Damage
- Police Involvement
- Assault
- Sexual Abuse
- Unsafe Sex
- Academic Problems
- Health Problems
- Suicide Attempts
- Death
WHAT DOES THIS HAVE TO DO WITH SOCIAL MEDIA?
EVERYONE IS USING IT

Especially college students
80% of College Students Use Twitter

46% of College Users Tweet Daily

63% OF College Users Check twitter at least once a day
Twitter lets anyone post what they’re feeling in 140 characters or less

- 500 million Tweets sent daily
- Tweets give us a firsthand account of someone’s life
- People are bound to use social media when they’re not sober
By analyzing student’s Twitter activity, we can predict when a user is under the influence, when they will drink next, and if they are considered to be at-risk.
Our main focus was current UMD students in the College Park, MD area

1550 Tweets were collected for initial analysis and classification
PREDICTION TOOL

- Our web tool analyzes a person’s Tweet and determines if they may be under the influence.
- If the tool thinks they are, it will Tweet at them and a friend of theirs, sending them a reminder to get home safely.
The first Tweets

- The 1550 Tweets initially collected not using the Twitter API were used to train an SVM that is used to predict other Tweets.
- These 1550 tweets were analyzed over their tone and content to see how users reacted to alcohol at the states of pre-drunk, drunk and hungover and which phrases were often used that showed problematic attitudes.
How The Tweets were gathered

The method
As stated before the Twitter API was not used. Instead the website tweetseeker was used. The website worked like Twitter advanced search but allowed us to save the resulting tweets.

The search conditions
- We limited our tweets to be from College Park and to be English.
- We searched for specific phrases not hashtags like drunk, hung over and shit faced.
Tweet Classification
We took all 1550 Tweets and assigned a number based on their sentiment:

- -1 for negative feelings
- 0 for neutral or having nothing to do with drinking
- 1 for positive feelings
FINDINGS

We took a chi squared test using the observed classification to the expected classification of -1 for each tweet and the probability returned from the test was 0.03 that all the tweets about alcohol we used was negative. As the percent was lower than 0.05, it means that the tweets about alcohol are much more varied in response than previous thought.
Trends in specific key words

Specific keywords that related to a specific state related to being drunk

- about 52% talking about the stress or depression from life
- 47% of these tweets talk about how much fun getting drunk is
- In comparison about 87% of the tweets when the users are drunk are very positive either enjoying the time as drunk or making jokes about their state
- nearly 93% of tweets when hungover were very negative about their current state
But overall there is a common trend of alcohol on Twitter, alcohol is usually seen as sign of trouble.

- Whether the tweet is positive or negative, usually the desire to drink is seen as a sign that their life is very hard or troubled.
- Even in the more positive posts being drunk is more of a sign of self-deprecation than a true enjoyment.
- Even posts that have pride in themselves for their drinking habits are more accepting that’s a flaw than anything else.
We asked 100 UMD students if they think the tool will be effective in reducing at-risk behavior, here’s what they said:

<table>
<thead>
<tr>
<th>Tweet Type</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet sent to self</td>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>Tweet sent to a designated friend</td>
<td>38</td>
<td>62</td>
</tr>
<tr>
<td>Tweet sent to both parties</td>
<td>60</td>
<td>40</td>
</tr>
</tbody>
</table>
CONCLUSION AND FUTURE APPLICATIONS

- Early detection of at risk drinkers who may develop alcoholism later in life
- Better understand drinking habits of college students to help lower at-risk behaviors while drinking
- Eventually have personalized profiles for all users with advanced NLP for better prediction
Gawker Inferencing
The untold story behind gossip data
Overview

1. Introduction - what is gawker
2. Acquiring the Dataset
   i. Downloading Articles
   ii. Scraping → JSON
   iii. Natural Language Processing Toolkit
3. Analysis / Insights Gained
   i. Machine Learning Classification
   ii. Link Graph
   iii. Tag graph
Introduction
The Dataset: Downloading Articles

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<url>
  <loc>http://gawkcr.com/sitemap_today.xml</loc>
  <lastmod>2015-05-03T19:59:16+04:00</lastmod>
</url>

<url>
  <loc>http://gawkcr.com/sitemap_news.xml</loc>
  <lastmod>2015-05-03T19:59:16+04:00</lastmod>
</url>

<url>
  <loc>http://gawkcr.com/sitemap_bydate.xml?start=2015-05-03T00:00:00&end=2015-05-09T23:59:59</loc>
  <lastmod>2015-05-09T23:59:59+04:00</lastmod>
</url>

<url>
  <loc>http://gawkcr.com/5-year-old-bruce-lee-fan-is-perfect-at-munchucks-and-ac-1701899755</loc>
  <lastmod>2015-05-03T17:30:00+04:00</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.7</priority>
</url>

<url>
  <lastmod>2015-05-03T16:56:00+04:00</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.7</priority>
</url>

<url>
  <loc>http://gawkcr.com/big-bird-actor-i-almost-died-on-the-challenger-and-i-c-1701893006</loc>
</url>
```

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<url>
  <loc>http://i.kinja-img.com/gawker-media/image/upload/s--kxs2FVwG--/c_fill,fi_progressive,q_80,w_636/1236411234973653029.jpg</loc>
</url>
```

Images:
The Dataset: Downloading Articles

Timeline of Page Download

Start

Downloading HTML/JS/CSS

AJAX Requests (viewcount)

Fully Loaded

CURL

WATIR

AJAX Query
The Dataset: Numbers to Note

1 script
100’s of threads
140,000+ articles.
14.7 billions bytes of HTML
The Dataset: Scraping/Parsing

Nokogiri

id
text
tags
date
blurb
author
num_pics
viewcount
comments
Analysis
Tag Graph
Valleywag
Link Graph
Machine Learning

- We want to find a correlation between an article’s features and its popularity (view count).
- Which article features actually influence popularity?
Article Features

- title
- author
- date article was published
- content
- tags
- number of pictures in article
Popularity

- An article’s popularity is measured by the number of views it has.
- Popularity is relative to the year in which it was published. An article with 20 views is popular in 2008 but very unpopular in 2013.
Did it work?

- No… at least not yet
- In classifying articles as either Unpopular, Average, or Popular:
  - Features: Title, Author, Tags, Pic_Count
  - Year: 2013, Number of Articles: 2000
  - Best f1 score: 0.51
- Tune: classifier type, parameters, features
Ali Eskandari
James Wills
Jake Fried
Capturing Company Stock Performance through Twitter Sentiment

By: Christi Capati, Kevin Chen, Kris Salvador, Brian Yedinak
Outline

I. Problem
II. Methods
  A. Twitter Dataset
  B. Yahoo Finance
  C. LingPipe
  D. Determining Correlation
III. Results
IV. Conclusions and Future Work
V. References
Problem

Does public sentiment affect a company’s stock performance?
- Focus is on social media specifically Twitter
- Analyze tweets directed at specific companies and determine sentiment
- Find correlation to stock performance for the respective companies

Example: When the new Samsung phone was released, what did users say about it on Twitter? Did stock success reflect users’ satisfaction/dissatisfaction?
Methods

Twitter Dataset

- Tweets targeted towards a company from Dec 18, 2013 through Feb 14, 2014
- Companies analyzed
  - AT&T
  - Verizon
  - T-Mobile
  - Sprint
  - Samsung
  - McDonald's
Methods

Yahoo Finance

- Retrieve stock performance for each company
- Stored in CSV file
- Example: AT&T Stock

<table>
<thead>
<tr>
<th>Date</th>
<th>Open</th>
<th>High</th>
<th>Low</th>
<th>Close</th>
<th>Volume</th>
<th>Adj Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/18/2014</td>
<td>33.23</td>
<td>33.23</td>
<td>32.63</td>
<td>32.82</td>
<td>22311900</td>
<td>30.68186</td>
</tr>
<tr>
<td>2/14/2014</td>
<td>33.38</td>
<td>33.38</td>
<td>33.06</td>
<td>33.15</td>
<td>18274800</td>
<td>30.99036</td>
</tr>
<tr>
<td>2/13/2014</td>
<td>32.83</td>
<td>33.5</td>
<td>32.71</td>
<td>33.49</td>
<td>28062200</td>
<td>31.30821</td>
</tr>
<tr>
<td>2/12/2014</td>
<td>32.91</td>
<td>33.07</td>
<td>32.81</td>
<td>32.94</td>
<td>16554700</td>
<td>30.79404</td>
</tr>
<tr>
<td>2/11/2014</td>
<td>32.62</td>
<td>33.04</td>
<td>32.51</td>
<td>32.95</td>
<td>26228800</td>
<td>30.80339</td>
</tr>
<tr>
<td>2/10/2014</td>
<td>32.45</td>
<td>32.46</td>
<td>32.22</td>
<td>32.44</td>
<td>20621000</td>
<td>30.32661</td>
</tr>
<tr>
<td>2/7/2014</td>
<td>32.18</td>
<td>32.36</td>
<td>31.95</td>
<td>32.3</td>
<td>29170600</td>
<td>30.19573</td>
</tr>
<tr>
<td>2/6/2014</td>
<td>32.31</td>
<td>32.31</td>
<td>31.74</td>
<td>32</td>
<td>45322100</td>
<td>29.91528</td>
</tr>
<tr>
<td>2/5/2014</td>
<td>32.4</td>
<td>32.43</td>
<td>32.04</td>
<td>32.08</td>
<td>29793500</td>
<td>29.99007</td>
</tr>
<tr>
<td>2/4/2014</td>
<td>32.08</td>
<td>32.45</td>
<td>32</td>
<td>32.45</td>
<td>35940100</td>
<td>30.33596</td>
</tr>
<tr>
<td>2/3/2014</td>
<td>33.32</td>
<td>33.32</td>
<td>31.9</td>
<td>31.95</td>
<td>68010500</td>
<td>29.86854</td>
</tr>
</tbody>
</table>
Methods

LingPipe
- Toolkit for processing text using computational linguistics
- Classify Tweets as Positive, Negative, and Neutral

JAVA app
- Parsed JSON data, extracting company directed tweets
- Ran LingPipe classifier on tweets
Methods

Determining Correlation

- Find Avg: \( \frac{\text{Total # of Pos Tweets}}{\text{Total # of Neg} + \text{Total # of Pos}} \)
- Graphed results using Excel Tools
## Results

### AT&T

<table>
<thead>
<tr>
<th>Date</th>
<th>Positivity Score</th>
<th>Number Pos</th>
<th>Number Neg</th>
<th>Number Neu</th>
<th>Total Pos</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu Jan 16</td>
<td>0.565217391</td>
<td>13</td>
<td>10</td>
<td>91</td>
<td>23</td>
<td>114</td>
</tr>
<tr>
<td>Wed Dec 18</td>
<td>0.615819209</td>
<td>109</td>
<td>68</td>
<td>627</td>
<td>177</td>
<td>804</td>
</tr>
<tr>
<td>Tue Jan 28</td>
<td>0.655688623</td>
<td>219</td>
<td>115</td>
<td>3247</td>
<td>334</td>
<td>3581</td>
</tr>
<tr>
<td>Fri Jan 17</td>
<td>0.65</td>
<td>234</td>
<td>126</td>
<td>999</td>
<td>360</td>
<td>1359</td>
</tr>
<tr>
<td>Sun Feb 09</td>
<td>0.595959596</td>
<td>59</td>
<td>40</td>
<td>330</td>
<td>99</td>
<td>429</td>
</tr>
<tr>
<td>Fri Feb 14</td>
<td>0.40952381</td>
<td>43</td>
<td>62</td>
<td>321</td>
<td>105</td>
<td>426</td>
</tr>
<tr>
<td>Mon Feb 10</td>
<td>0.62248996</td>
<td>155</td>
<td>94</td>
<td>1113</td>
<td>249</td>
<td>1362</td>
</tr>
<tr>
<td>Thu Dec 19</td>
<td>0.6666666667</td>
<td>106</td>
<td>53</td>
<td>437</td>
<td>159</td>
<td>596</td>
</tr>
<tr>
<td>Fri Jan 03</td>
<td>0.596491228</td>
<td>68</td>
<td>46</td>
<td>728</td>
<td>114</td>
<td>842</td>
</tr>
<tr>
<td>Wed Jan 29</td>
<td>0.714285714</td>
<td>40</td>
<td>16</td>
<td>467</td>
<td>56</td>
<td>523</td>
</tr>
<tr>
<td>Sat Jan 04</td>
<td>0.786641929</td>
<td>424</td>
<td>115</td>
<td>2504</td>
<td>539</td>
<td>3043</td>
</tr>
</tbody>
</table>
Conclusions & Future Work

- Larger dataset with Tweets from more days in the year
- More time to build a better classification model
- Larger machine
  - ~40gbs of JSON data to parse and classify
- Classifier
  - Content and context
    - Sprint (the verb) vs. Sprint (the company)
  - Emoji
  - Retweets
  - Slang (e.g. booty, h8, whack, mfs, etc.) & foreign language tweets
  - Spam tweets ("Enter to win a Samsung Galaxy...")
References

CMSC498J: Final Project
Churniness Effects of Tweets on Neighbors

G8: Dylan Zingler, Gilad Kempenich, Kevin Gutierrez, Zachery Knopp
Imagine these people as your friends/followers on a social network.

What would their text look like? Lots of emotion, ALL CAPS, !!!!!!, profanity (a$$, etc.)
For large companies and brands it is important to understand how one user's tweets may or may not affect the nodes or users it is directly connected with.

Our project researches how one user’s positive or negative input through tweets influences the positive or negative input of their neighbors' tweets.

In Twitter’s social network, users are able to subscribe or follow other users and see their contributions to the network, mainly in the form of tweets and twitpics.
Problem

It is important for many purposes, including cascading effects, which can be systematically interrupted, reversed or enhanced, by a brand in order to influence and shape the network to their favor.

The scope of our problem will be limited to the data that we access to. This includes a group of labeled tweets from phone companies; Verizon, AT&T, and T-Mobile.

Tweets were already labeled Non-Churny or Churny. We also had mass unlabeled tweet data that was process and classified into Non-Churny or Churny.
Loyalty Factors

Main factors that contribute to a customer’s loyalty:

- Price
- Competition
- Poor or inferior brand service
- Customer Service
- Bad Anticipation of Customers needs
A new factor is emerging, a person's loyalty to a brand might be influenced by the social media they are exposed to and the peers they interact with on these networks.

Quote from “Social Media and Prosumerism”: Through the exchange of information, members of social networking communities work together to achieve better products and services in a manner that puts power in the hands of the consumer who becomes an advocate and voice for a brand. (Buzzetto-More 2015)
Predictions/Hypothesis

When a user subscribes to another user, they read and have access the others information more readily than if there was no connection.

Since users see each others information, they develop opinions and relay it through their own input to the social network.

We believe a users churniness level will correlate with their neighbors behavior.
Twitter telecom tweets
- Three Datasets (AT&T, Verizon, and T-Mobile)

Example Line:
530707988,false,,finalized,3,8/20/2014 14:35:55,Churny,1.0,at&t,,Wed Oct 09 18:49:23 +0000 2013,I can't wait to switch to tmobile. Sick of at&t stupid a$$.,388013339764613121,185102676

Twitter large span tweets
- Large number of tweets from a few months in 2013
- Stored in .rar files, only used a selection

Twitter social graph
- A file containing their a labeled users neighbors (Based on UUID’s)

Example Line:
47141288 14135350-33230773-419928486-284991184-15931637-84061721-23508439-...continues.
Churny vs Non-churny Classifier

Naive Bayesian Classifier

Uses N-grams features to determine churniness

- unigrams: "I", "am", "leaving", "verzion"
- bigrams: "I am", "am leaving", "leaving verzion" =>
  {"I", "am", "leaving", "verzion", "I am", "am leaving", "leaving verzion"}
Analyzing Datasets

Use already processed tweets to create a Bayesian Classifier

Process set of raw tweets and store user churny status

Use social graph to connect users and draw conclusions
Plotting data

Full graph of subset
Plotting data

Plotting Churny users only for subset

- = Churny
- = Non-churny
Churny vs Non-churny user

Non-churny user neighbors

Churny user neighbors

Churny

Non-churny
Results

Found that churny users were connected throughout subset of Social Graph.

Churny users have more churny neighbors than Non-churny users.

Reasons for difficulty:
- There are not many Churny tweets so it becomes hard to identify a correlation through the large number of Non-churny tweets.
• We found that the user’s churny behavior is correlated with the churny behavior of the user’s neighbors

• The data we collected that as more of a user’s neighbors make a churny tweet about a brand that user is more likely to make a churny tweet too

• We attributed homophily as the reason for the correlation between churn and friends’ churn
Cascading

Information cascades are the main idea behind the “following the crowd” mentality.

Any individual has a choice to make whether they will positively or negatively tweet about a company.

The probability that a user will produce negative or positive data about a brand is calculated using the previous user information and is modeled off of Bayes Theorem.

\[
P(A|\text{Previous, Personal signal}) = \frac{pq^a(1-q)^b}{pq^a(1-q)^b + (1-p)(1-q)^aq^b}
\]
Twitter as a Mental Illness Diagnosis Tool: Exploring user commonalities

Members: Henry Zhou, Walid Ali, Jordan Clarke, Justin Le
Introduction

Barriers:

- issues are seldom discussed
- difficult to diagnose
- require scheduling physician/specialist visits that are not done due to personal/financial reasons
Application of data
Approach/Mission

MetaMind Sentiment analysis

Classifier

Twitter Feed

Contextual Data
Methodology - Classifier

- NLP (Natural Language Processing) and machine learning -> linear regression
- Metamind processing training data
  - 100 positive & 100 negative sentences from Reddit
- Depression classifier - positive (no sign of depression) & negative
  - 85% accurate
Methodology - Twitter Scraper

- **APIs**
  - **Twython** - Python wrapper for Twitter REST API
  - **Metamind**
  - REST-based data scraping

- **Target affiliation network of interest** -> 200 followers
- **Run all posts against classifier to determine sentiment**
- **0.75+ score threshold for users**
- **Retain users with 35+ posts**
Findings

@Fitspirational
1.52M followers

@DiabetesMine
22.6K followers

@YoungEnt
370K followers

@yogadistrict
3,321 followers
Anonymous

Depressed - anorexic - alcohol & drugs - I want to help everyone but I can't make the pain away - addicted - bullied.

Anomalous

Just another depressed teenager... No big deal.. | 16 | zero days clean | secret account | I just let my thoughts eat me | sorry for everything I've done...

canada
Further Explorations & The Next Step

- Current Classifier Limitations
  - Coppersmith’s Classifier Access
- Real-world Applications/Integration- Happsee by Vikas Paruchari
- Twitter API Application Rate Limitations
  - Usage of REST API (requests) allowed for ease of data exploration, but...
  - Maintaining long-term stream (Public Stream & User Stream) can gather much more
References


CMSC773 Reddit Data - Philip Resnik