

## Social Media Analytics: Data Mining Applied to Insurance Twitter Posts

CAS Ratemaking and Product Management Seminar  
Roosevelt C. Mosley, Jr., FCAS, MAAA  
Pinnacle Actuarial Resources, Inc.  
March 21, 2012

Experience the Pinnacle Difference!

---

---

---

---

---

---

---

---

## Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

---

---

---

---

---

---

---

---

## Social Media Analytics

- The growth in social media
- Background on Twitter
- Data
- General descriptive statistics
- Processing the data
- Analysis – identifying the themes
- Analysis challenges
- Application of social media analytics

---

---

---

---

---

---

---

---

**FINNACLE**  
The Center of Choice

## The Growth in Social Media



---

---

---

---

---

---

---

---

## Social Media Defined

- **Social media:** a group of Internet-based applications that build on the ideological and technological foundations of [Web 2.0](#), and that allow the creation and exchange of [user-generated content](#)
- Building blocks
  - Identity
  - Conversations
  - Sharing
  - Presence
  - Relationships
  - Reputation
  - Groups

Kaplan, Andreas M.; Michael Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media". *Business Horizons*

---

---

---

---


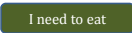


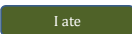
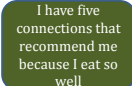

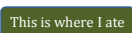

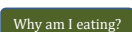

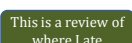


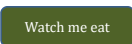
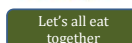
---

---

---

---

## Social Media Platforms

---

---

---

---

---

---

---

---

### Social Media – Explosive Growth

- Facebook has 750 million users
- 30 billion pieces of content is shared on Facebook every month
- As of May 2011, there were on average 190 million tweets per day
- Google+ reached 10 million users in 16 days
- People upload 3,000 images to Flickr every minute

Source: <http://www.jeffbullas.com/2011/09/02/20-stunning-social-media-statistics/>

---

---

---

---

---

---

---

---

### Business Has Taken Notice

- Two-thirds of comScore’s U.S. Top 100 websites and half of comScore’s Global Top 100 websites have integrated with Facebook
- Many businesses now have established Twitter accounts in an attempt to connect with current and potential customers
- 80% of companies use LinkedIn as a recruitment tool
- Companies spent over \$3 billion to advertise on social media sites in 2011, an increase of 55% over 2010

---

---

---

---

---

---

---

---

### Insurance Facebook Fans

Page	Category	Fan Count (December, 2011)	Percentage Growth
Flo, The Progressive Girl	Mascot	3,336,486	1.5
Farmers Insurance	Corporate	2,360,972	-0.5
State Farm Nation	Demographic	1,353,524	1.1
Mayhem	Mascot	1,129,941	1.7
AFLAC Duck	Mascot	293,496	0.6
USAA	Corporate	208,732	1.6
The Gecko	Mascot	204,593	1.8
GEICO	Corporate	202,825	2.2
State Farm Insurance	Corporate	193,864	8.1
New York Life	Corporate	154,390	19.6

Source: Customer Respect Group. "Social Eyes: The Insurer's View of Social Media."

---

---

---

---

---

---

---

---

## Insurance Twitter Followers

Company	Category	Followers (December, 2011)	Percent Change
State Farm Nation	Corporate	28,218	0
Allstate Insurance	Corporate	25,884	1
USAA	Corporate	23,742	4
New York Life	Corporate	23,344	31
State Farm Insurance	Corporate	18,856	5
VPI	Pet	17,551	7
Hartford Achieve	Advocacy	17,524	-2
AFLAC Duck	Mascot	13,709	2
The Hartford	Corporate	10,913	2
Progressive Insurance	Corporate	10,531	3

---

---

---

---

---

---

---

---

---

---

## Are You Taking Advantage of Social Media?

- Insurance companies are investing significant resources in a social media presence
- Current and potential customers are voluntarily sharing intimate details of their life with the world
- Current and potential customers are interacting with companies on a very personal level
- This information can be applied in different ways (service, marketing, competitive monitoring)

---

---

---

---

---

---

---

---

---

---



## Twitter Background

---

---

---

---

---


---

---

---

---

---



**Data**

---

---

---

---

---

---

---

---

**Data Used for Paper – Dataset #1**

- Tweets including #allstate – 68,370
- Dates: July 29, 2010 – August 12, 2011
- Downloaded from twapperkeeper.com – no longer exists
- Data
  - **user**: the username that sent the tweet
  - **tweet**: the content of the tweet
  - **timestamp**: the date and time the tweet was sent (GMT)
  - **tweet ID**: Twitter identification number of the tweet
  - **geo**: latitude and longitude of the user

---

---

---

---

---

---

---

---

**Data Used for Updated Analysis – Dataset #2**

- Keyword searches for State Farm, Allstate, Geico, esurance, and #Progressive – 176,694 tweets
- January 25 – February 12, 2012
- Tracked through hootsuite.com
- Data
  - **text**: content of the tweet
  - **to user id**: specific tweet recipient
  - **from user**: sender of the tweet
  - **iso language code**: language of tweet
  - **source**: where did the tweet originate?
  - **profile image**: picture of user
  - **geo**: latitude and longitude of the user
  - Date and time

---

---

---

---

---

---

---

---

### Sources of Social Media Data

- Third party data aggregators (hootsuite, GNIP)
- API
- Company developers
- Screen scraping

---

---

---

---

---

---

---

---



### General Descriptive Statistics

---

---

---

---

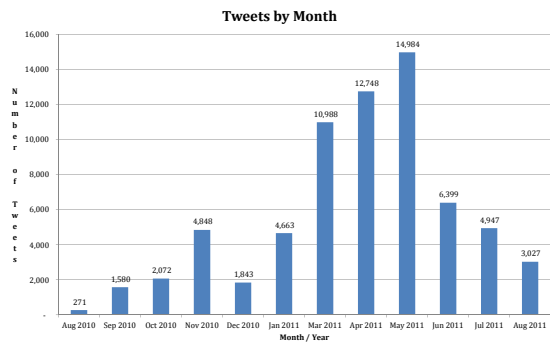
---

---

---

---

### Tweets Per Month - #allstate



---

---

---

---

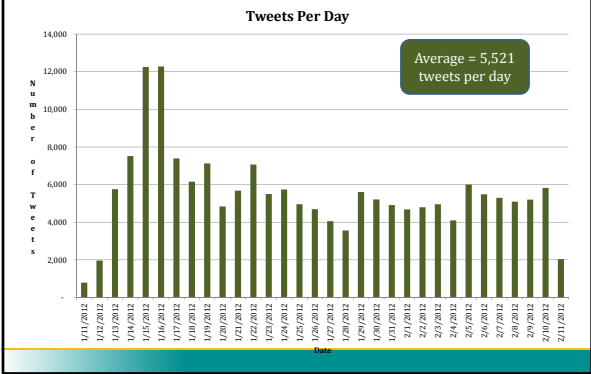
---

---

---

---

## Tweets per Day - Dataset #2



---

---

---

---

---

---

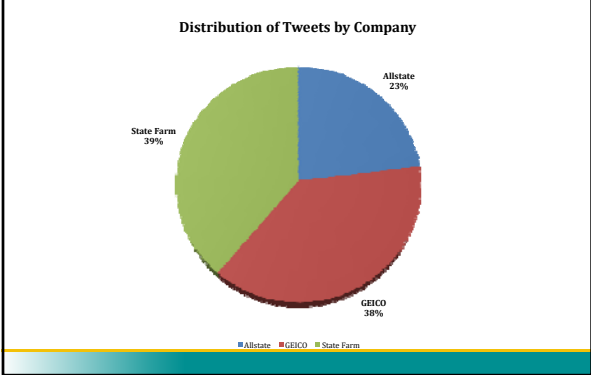
---

---

---

---

## Distribution of Tweets by Company



---

---

---

---

---

---

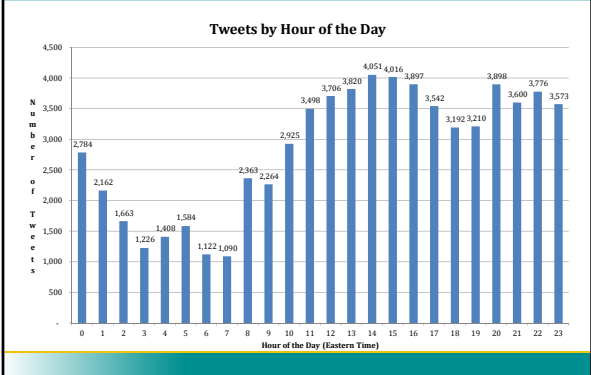
---

---

---

---

## Tweets per Hour - Dataset #1



---

---

---

---

---

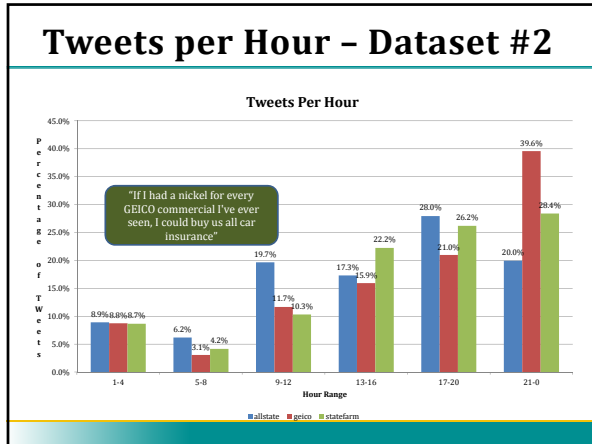
---

---

---

---

---




---

---

---

---

---

---

---

---

---

---

---

---

## Data Processing Steps

---

---

---

---

---

---

---

---

---

---

---

---

## Data Processing Steps

- Remove punctuation and symbols (retain @ and #)
- Parse the tweet (35 words worked for Twitter – will need many more for other sources)
- Change table structures from tweets in rows to tweets in columns – keep indicator of order
- Correct spelling errors
- Add word indicators

Tweet ID	User	Tweet	Word1	Word2	...	Word35
1	@mosley	Text of tweet	W1	W2	...	W35

Tweet ID	Word Order	Word
1	1	Word1
1	2	Word2
...	...	...
1	35	Word35

---

---

---

---

---

---

---

---

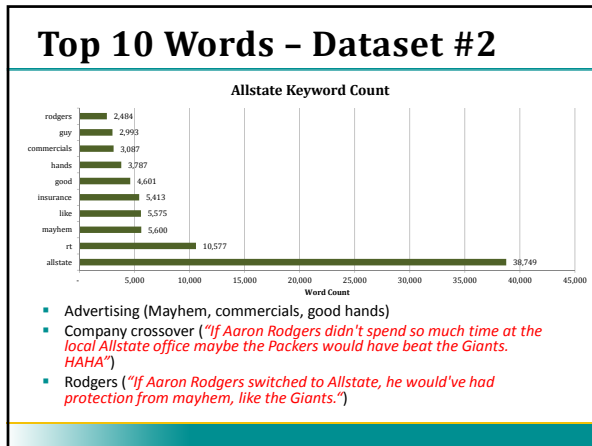
---

---

---

---






---

---

---

---

---

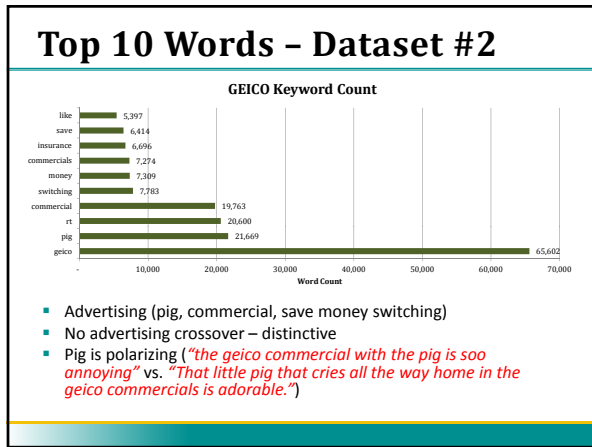
---

---

---

---

---




---

---

---

---

---

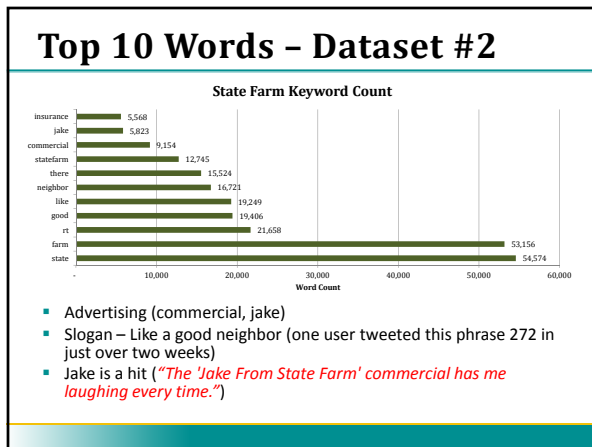
---

---

---

---

---




---

---

---

---

---

---

---

---

---

---

## Spelling Errors

- **Levenshtein Edit Distance (LED):** number of insertions, deletions, or replacements of single characters that are required to convert one string to another.
- **Generalized Edit Distance (GED):** minimum cost sequence of operations for constructing string2 from string1
- Smaller distances indicate similar words
- Words can be edited and corrected based on this analysis

---

---

---

---

---

---

---

---

## Word Indicators

Tweet ID	User	Tweet	allstate	insurance	commercial	company	...
19	@rmsosky_pur	Allstate insurance company	1	1	0	1	...

- Add indicators to table to indicate which keywords are present in each tweet
- 116 keywords in dataset #1, 154 keywords in dataset #2
- Indicators are adjusted for misspellings and tenses

---

---

---

---

---

---

---

---



## Analysis – Identifying Themes in the Data

---

---

---

---

---

---

---

---

### Correlation Analysis - Dataset #1

Number	var1	var2	Cramer's V Statistic
1	state	farm	0.861
2	financial	personal	0.803
3	good	hands	0.734
4	agency	purchase	0.683
5	jobs	grave	0.661
6	esurance	answer	0.612
7	girl	neighbor	0.508
8	youtube	jonas	0.505
9	watch	neighbor	0.489
10	work	neighbor	0.483
11	Love	basketball	0.454
12	Youtube	video	0.452
13	Geico	progressive	0.427
14	Girl	watch	0.420
15	insurance	company	0.413
16	Farm	neighbor	0.405
17	Agent	exclusive	0.394
18	Girl	work	0.387
19	Watch	work	0.366
20	Billion	answer	0.360

- Only compares pairs of words
- No real understanding of how many tweets are impacted

---

---

---

---

---

---

---

---

---

---

---

---

### Clustering/Segmentation

- Unsupervised classification technique
- Groups data into set of discrete clusters or contiguous groups of cases
- Performs disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative input variables and cluster seeds
- Data points are grouped based on the distances from the seed values
- Objects in each cluster tend to be similar, objects in different clusters tend to be dissimilar

---

---

---

---

---

---

---

---

---

---

---

---

### Cluster Lift - Dataset #1

$$\text{Cluster Lift (word)} = \frac{\text{Percentage of tweets in a cluster that include word}}{\text{Percentage of all tweets that include word}}$$

Cluster	insurance	rt	commercials	arena	good	mayhem	job	like
1	0.519	7.879	1.778	0.000	0.741	2.333	0.000	17.635
2	0.300	0.082	0.000	0.000	0.000	0.000	6.631	0.010
3	1.471	0.572	0.000	0.000	0.000	0.000	0.000	0.000
4	0.002	7.879	0.598	0.000	0.000	0.399	0.325	0.000
5	0.169	1.697	2.118	0.018	0.047	0.149	0.055	1.287
6	0.000	0.000	0.025	0.000	0.000	0.000	0.315	0.000
7	0.313	0.000	1.335	0.063	0.047	1.990	0.081	17.635
8	0.000	0.000	0.000	0.000	0.000	0.000	0.055	0.000
9	5.215	0.408	0.000	0.000	0.222	0.000	0.641	0.057
10	5.196	0.484	0.014	0.000	0.036	0.000	1.555	0.065

---

---

---

---

---

---

---

---

---

---

---

---

## Keywords - Dataset #2

Cluster #	Number of Tweets	Distribution	Category	Keywords
5	1,221	0.7%	commercial	jay jake wearing sounds hideous
6	42	0.0%	commercial	commercial check game best hilarious still show raj
8	74,208	42.0%		
11	1,340	0.8%		auto company year progressive
13	911	0.5%	commercial	pig piggy little weee weeeee makes weeeee
15	135	0.1%	commercial	rodgers aaron about best ad dude lebron thing
19	1,127	0.6%	commercial	allstate good rodgers aaron hands should michaelstrahan leave funniest wow read
20	4,817	2.7%	employment	agent jobs job auto geico company
21	42	0.0%	employment	no about office life better still call show people own
23	985	0.6%	commercial	time more packers giants hanging spent watching checking less discounts offices film
24	1,823	1.0%	commercial	best weeee weeeee
31	1,241	0.7%	parody	allstate like rodgers mayhem aaron giants switched protection espn_colin
32	1,051	0.6%	agency	job great own
33	1,074	0.6%	advertising	new piggy ad weeee raj youtube
34	1,991	0.9%	advertising	rodgers aaron packers discount check double
35	985	0.6%	advertising	rodgers aaron check double hilarious show raj
36	1,908	1.1%	advertising	progressive call

Opening an Allstate Agency is a great way to take control of your financial future. #insurance

---

---

---

---

---

---

---

---

---

---

---

---

## Key Themes - Dataset #1

Theme	Number of Tweets	Percentage of Tweets
advertising	12,976	18.7%
agency	4,150	6.0%
arena	5,621	8.1%
blank	21,002	30.3%
claims	1,466	2.1%
competition	2,467	3.6%
description	5,499	7.9%
employment	2,327	3.4%
foundation	957	1.4%
news	662	1.0%
other	6,740	9.7%
praise	1,464	2.1%
quotes	1,807	2.6%
roadside	1,232	1.8%

---

---

---

---

---

---

---

---

---

---

---

---

## Association Analysis

- Background in market basket analysis
- Identification of items that occur together in the same record
- Produces event occurrence as well as confidence interval around the occurrence likelihood
- Can lead to sequence analysis as well, which considers timing and ordering of events

---

---

---

---

---

---

---

---

---

---

---

---

## Association Analysis Formulas

$$\text{Support} = \frac{\text{Transactions that contain items A \& B}}{\text{All transactions}}$$

$$\text{Confidence} = \frac{\text{Transactions that contain items A \& B}}{\text{Transactions that contain item A}}$$

$$\text{Expected Confidence} = \frac{\text{Transactions that contain item B}}{\text{All transactions}}$$

Transactions = tweets  
Items = words

---

---

---

---

---

---

---

---

---

---

---

---

## Association Rules - Dataset #1

Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction		Left Hand of Rule	Right Hand of Rule	Rule Index
				Count	Rule			
0.632	73.391	0.618	116.070	422	neighbor ==> there & like & good	neighbor	there & like & good	73
0.754	86.831	0.618	115.194	422	neighbor & like ==> there & good	neighbor & like	there & good	76
0.754	81.043	0.682	107.516	466	neighbor ==> there & good	neighbor	there & good	115
0.771	79.174	0.618	102.644	422	neighbor & good ==> there & like	neighbor & good	there & like	124
0.888	91.053	0.760	102.486	519	jonas ==> xthetxt	jonas	xthetxt	135
0.834	85.502	0.760	102.486	519	xthetxt ==> jonas	xthetxt	jonas	136
0.771	75.130	0.632	97.402	432	neighbor ==> there & like	neighbor	there & like	177
1.762	99.167	0.697	56.273	476	financial & answer ==> esurance	financial & answer	esurance	277
0.921	50.911	0.777	55.299	531	tv ==> mayhem & ad	tv	mayhem & ad	283
1.527	84.420	0.777	55.299	531	mayhem & ad ==> tv	mayhem & ad	tv	282
0.700	37.679	0.618	53.855	422	like & good ==> there & neighbor	like & good	there & neighbor	295
1.639	86.285	0.618	53.855	422	there & neighbor ==> like & good	there & neighbor	like & good	296
1.041	53.679	0.777	50.620	531	ad ==> tv & mayhem	ad	tv & mayhem	315
1.762	88.516	0.733	50.229	501	answer ==> esurance	answer	esurance	323
1.639	81.913	0.689	49.968	471	neighbor ==> like & good	neighbor	like & good	333
1.399	65.843	0.897	47.056	613	video ==> youtube	video	youtube	364
2.122	87.639	0.809	41.294	553	state & car ==> farm	state & car	farm	530
0.924	38.138	0.809	41.294	553	farm ==> state & car	farm	state & car	529
2.122	80.923	0.950	38.130	649	state & insurance ==> farm	state & insurance	farm	588
1.174	44.759	0.950	38.130	649	farm ==> state & insurance	farm	state & insurance	587

---

---

---

---

---

---

---

---

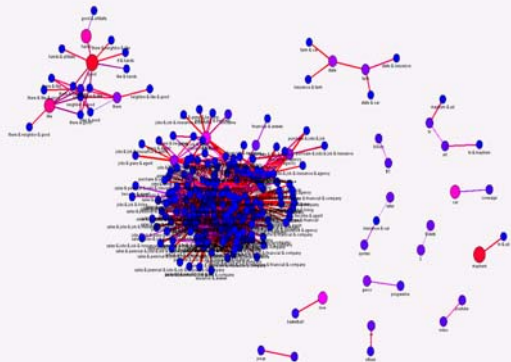
---

---

---

---

Association Analysis  
Link Graph




---

---

---

---

---

---

---

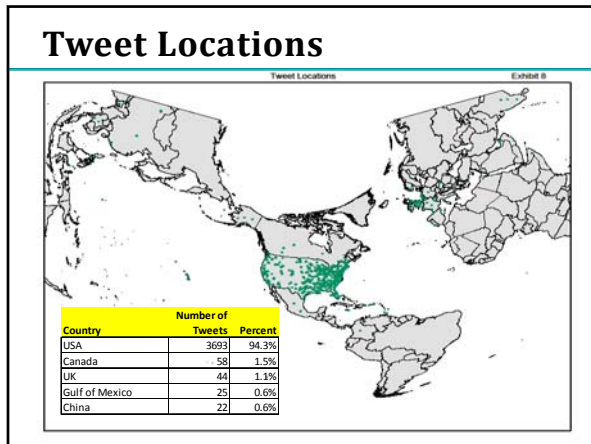
---

---

---

---

---




---

---

---

---

---

---

---

---

- ### Analysis Challenges
- Accessing and collecting information
  - Context
  - Relevance
  - Influence
  - Sentiment
  - Raw, unfiltered customer data
  - Near real-time analysis needed

---

---

---

---

---

---

---

---

- ### Applications of Social Media Analytics
- Customer service
  - Understanding customer sentiment relating to company (advertising, etc.)
  - Competitive intelligence
  - Broad market trends

---

---

---

---

---

---

---

---



---

## Social Media Analytics

Visit us at [www.pinnacleactuaries.com](http://www.pinnacleactuaries.com)

Roosevelt C. Mosley, Jr.  
309.807.2330  
[rmosley@pinnacleactuaries.com](mailto:rmosley@pinnacleactuaries.com)  
Follow me on Twitter: [rmosley@par](#)  
Connect with me on LinkedIn and Google+

---

Experience the Pinnacle Difference!

---

---

---

---

---

---

---

---