Spring 2013

USING SOCIAL NETWORKS FOR ASSESSING COMPANY SALES AND MARKETING PROGRAMS

Vance Tomchalk
San Jose State University

USING SOCIAL NETWORKS FOR ASSESSING COMPANY SALES AND MARKETING PROGRAMS

A Paper

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Computer Science

by

Vance Tomchalk

December 2012

The Designated Project Committee Approves the Paper Titled

USING SOCIAL NETWORKS FOR
ASSESSING COMPANY SALES
AND
MARKETING PROGRAMS

by

Vance Tomchalk

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

December 2012

Dr. Teng Moh	Department of Computer Science
Dr. Mark Stamp	Department of Computer Science
Alex Rivlin	Project Committee Member

ABSTRACT

USING SOCIAL NETWORKS FOR
ASSESSING COMPANY SALES
AND
MARKETING PROGRAMS

By Vance Tomchalk

During the course of an extended sales period for a company's given product line, there are many events that affect the success of its sales. Some of these events include economic downturns, unforeseen shortages and delays that effect the supply chain for the product, and product quality issues that change the perception of the product as a safe and cost-effective choice. In many instances, these events can be tracked by analyzing the signals and messaging present in the social networking media. This analysis requires careful consideration, which the metrics provided by software tools and algorithms lend considerable aid.

TABLE OF CONTENTS

LIST of FIGURES

LIST of TABLES

1. BACKGROUND

Among the different ways to regard a social medium and its marketing include viewing it as a collective: the aggregate signal that represents a bundle of communications, messaging, and postings. It may be analogous to a football game or soccer match. When the hometown crowd cheers, it may cheer collectively for the overall success of the team, but the crowd may be composed of different sections that may be cheering for different reasons. For example, based on their proximity to the playing field, some may cheer in response to their action in front of them, their favorite player, or they may even be cheering in response to a team cheerleader (and perhaps a self-appointed cheerleader or rally-mascot in the stands). But whatever their individual or group reason -- the cheering (or perhaps booing) -- can be experienced collectively.

To carry the analogy a bit further, a cheering stadium is an indication of fan sentiment, but the fan sentiment is often amplified by commercial prompts courtesy of the team franchise and its sponsors. In short, fan sentiment may exist without this type of amplification , but the degree to which a fan cheers and shouts, how long they cheer and shout, and whether they return to the stadium another day, is in large a function of how well the team franchise and sponsors create a positive *cheering* effect for the team. The overall fan-base is self-amplifying as well , in the sense that cheering over a goal or touchdown scored reinvests its own energy in more cheering, until, of course, the fans are quite hoarse.

Giving this sports analogy a last pass, negative sentiment comes to note. It has the same characteristics as cheering, but in the opposite direction. Negative fan sentiment

rarely overshadows cheering fans, but is noteworthy when the rest of the cheering stadium is silent. The reason for this analogy is to understand a popular social medium, such as Twitter, in more literal terms. A series of dense, concentrated messages is like a stadium full of noise over a given time span. Within that time, the tweeters, which can be comprised of casual web users to passionate consumers to marketers, espouse their interests in a forum which is, in a sense, a virtual stadium of cheering and booing for a given topic. The extent to which these tweets affect the failure or success of the subject at hand, whether it is a sports team – or a business – is the focus of this paper.

2. EVALUATING SOCIAL MEDIA AND BUSINESS

As indicated in the previous section, social media has the effect of cheering, commenting, advising and even booing in the virtual stadium known as the internet. In order to evaluate this effect in business terms, this paper will utilize a method for drawing on the content provided by Twitter and use that content to measure how Twitter messages coincide and perhaps forecast sales from a popular car company, Toyota.

2.1 Related, Current and New Work

The Institute of Business and Forecasting and Planning ("Institute of Business and Forecasting," 2012) is one of the premiere business communities for analyzing and compiling the latest trends in business supply chain forecasting and planning. Although it has frequent seminars, tutorials and other activities, when searching its associated web site, the keyword combinations of *social network sales, social media* sales and *data mining sales* do not return any relevant results in the context of sales analysis and

forecasting and social networks. Applied to searching conferences and proceedings, the key terms did yield did find one work that addressed time series analysis of sales data with respect to moving averages (Geng, 2010), but without relevant reference to applying social networking data mining to sales assessment and forecasting. Other work included detailed analysis of vehicle sales trends (Shahid, 2009) but without significant reference to mining social networking data and any deeper analysis of its data in terms of a supervised or unsupervised approach. Other work has involved using the web as source of business decision support (Zhang, 2010), but in that instance, the data was analyzed in terms of how social media "controversy" was a catalyst and predictor of sales, rather than analyzing all of the terms in a more unsupervised fashion.

More recent work, within the last few months, coming in parallel with the proposed idea for this paper, includes extracting various terms across diverse social media regarding vehicle quality issues ("Mining media to find vehicle defects", 2012). The methods described in that referenced paper will be published within the same month as the date of this paper, but are not specifically focused on evaluating all extracted terms, which this paper does, but rather, those terms that pertain to vehicle quality. Also, that approach does not appear to focus on sales months in a time series approach, while this paper explicitly makes that comparison. But it is interesting that both the approach of that referenced paper and the methods described in this paper allude to one central theme: business decision making tools are in the process of changing, with the advent of social data mining and analysis leading the way.

Ultimately, this paper is in concert with the trend recognized by a recent Forbes magazine article (Holmes, 2012) as noted in the following excerpt.

> At a glance, directors and department heads can see real-time analysis of social metrics and use this to inform business decisions. These tools are already being used by Nestle to track customer sentiment, GE to speed up repairs to the electrical grid, the auto industry to predict recalls, Wall Street to forecast stock prices and T-Mobile to prevent customer defections.

As such, it is exciting and reasonable to expect that everyday sales metrics, including actual sales, billings, and forecasts, will rely more and more on analyzing how social networking data analysis adds another layer of meaning to that data, thus enhancing business decision support and giving companies a competitive edge.

2.2 Approach and Outline

Essentially, Sec. 3 and 4 will focus on the methods of obtaining Twitter content and corresponding Toyota content; Sec. 5 will discuss the value of obtaining a term document matrix as a first step in Principal Component Analysis, and Sec. 6 for cluster analysis, Sec. 7 for the Conclusion, and Sec. 8 for Further Directions.

3. TOYOTA BACKGROUND AND MARKETING PROGRAM

This section will give a brief overview of Toyota sales and its marketing efforts.

3.1 Toyota Sales

As a car company, Toyota covers a variety of vehicle classes, each comprised of models that afford tradeoffs for luxury, price, durability and economy, among other features. In terms of sales, both volume and cash, Toyota is top tier auto manufacturer. Breaking it

down further, Toyota's main car divisions include economy/mid-range, Lexus, SUV, pickup and Lexus SUV's. Having this level of coverage in the automobile industry requires Toyota to offer a number of incentive programs and accompanying advertising, in order to remain competitive. One relatively currently selling model is the Toyota Prius, which offers outstanding mpg, improved safety features, smooth handling, and competes well with other make-and-model, such as the Honda Civic hybrid. Taking all of these features into consideration, there is plenty for Toyota to promote, particularly in the current business cycles where fuel prices are particularly high and long-term buying considerations factor heavily in a car's resale value.

In addition to relative new-comers like the Prius, Toyota has retained a number of other big-sellers, such as the Camry, Corolla, Rav4, and Tacoma. Although some may consider these vehicles as *legacy*, they are replete with new and improved features and Toyota, in line with other auto companies, is constantly marketing how they have kept improving them, year-after-year. Both the Camry and the Prius have fared well in 2012 according to Toyota's Pressroom ("Toyota USA Newsroom," 2012), an engineering and press-release portal that covers many facets of the organization, including sales, fact-sheets, recall notifications, marketing, product offerings – and social media releases.

3.2 Toyota Marketing Programs

One of the more popular buyer incentive programs is the Shareathon program ("Toyota Announces Second Annual Shareathon Program", December 2011). In this program, Toyota sets aside a dedicated web site for existing or potential customers to register and effectively enlist themselves not only as customers but also as promoters of Toyota

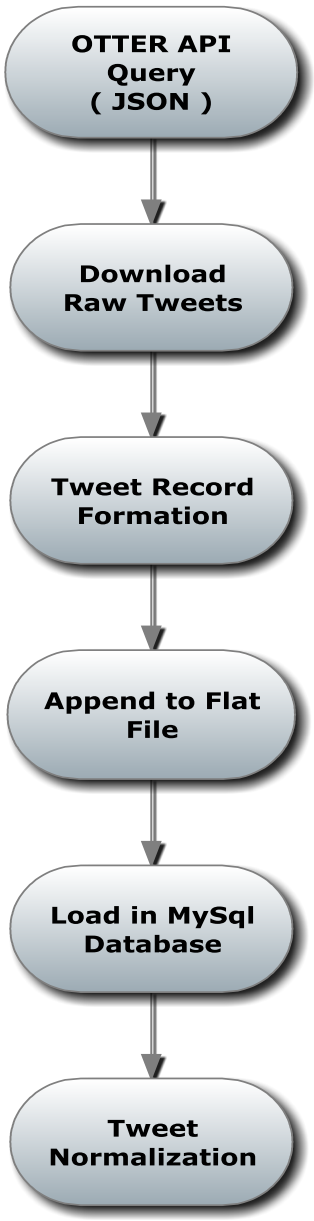
products. In order to benefit, registrants must tweet a scripted message regarding Toyota. In a note of irony, the Shareathon opened the certificate of eligibility to the first one hundred forty individuals, a deliberate reference to the Twitter one hundred forty character limit on tweets. Those who qualify would receive a $500 minimum, with a $50 bonus per retweet, up to a $1000 maximum.

The December 2011 Shareathon also prepared publicly available "interactive visualization" so participants could view the cluster of tweets and retweets associated with their given tweet. Overall, Toyota's use of the Shareathon program definitely generated significant social network traffic to elevate consumer interest in their products. In subsequent sections of this paper, the Toyota tweets, including those attributed to Shareathon, will be analyzed in the context of how they boosted consumer interest.

4. TWITTER SOCIAL NETWORK

Similar to other social networks, Twitter requires account registration and allows the registrant to broadcast and receive messages from registered relationships throughout the Twitter network. These broadcast messages are simply referred to as *tweets.* This section will describe different methods for extracting and collecting tweets from a programming and data mining perspective. The following is an overview of how tweets are gathered and loaded.

Figure 4-1


OTTER API
Query
(JSON)
Download
Raw Tweets
Tweet Record
Formation
Append to Flat
File
Load in MySql
Database
Tweet
Normalization

4.1 API's For Twitter Message Extraction

There are a number of different types of API's to retrieve Twitter message. Twitter itself provides two different architectures for retrieving twitter content: a streaming API and an API that retrieves twitter messages that have been archived. The streaming API ("The Streaming APIs ", 2012) as its name indicates, retrieves tweets real-time, through a stream capture object, which can be implemented through almost any language, such as perl , Java, .NET, and python,. One real advantage of the streaming API is that it will allow viewing of trending messages on the web as they becomes available through Twitter. The streaming API requires a secure key, based on OAuth, and is subject to certain use requirements. One significant drawback is when the streaming API breaks, either through hardware or software issues.

Twitter also offers an API that does not require an OAuth key and can be readily implemented through any scripting language, such as perl or python. This API ("Using the Twitter Search API", 2012) provides the ability to obtain fifteen hundred tweets per day per query, going back seven days. Although this API provides the ability to backfill when the real-time API goes offline, it does not provide the ability to return content that would allow one to evaluate trends that go months or even years back.

In any business or economic field, data needs be collected over many months or perhaps years, to understand month-to-month and year-to-year sales patterns. The demand for this type of data mining archival services ultimately lead to the advent of large server farms that harvest Twitter messages, collecting these messages as they become available, and storing them indefinitely. One such company that provides this

type of service is Topsy, a company which provides an interface that is very similar to Twitter, allowing the developer to use query terms, time ranges, return format, and language specification. In its terms of service, Topsy provides a keyless API that is limited to three thousand query requests per day. The limitation is enforced by the machine IP that is broadcast to the server with each call made to the Topsy repository. It also provides the opportunity to evaluate and eventually purchase a key that allows for up to seven thousand requests per day. The evaluation key is valid for one month, after which a decision to purchase must be made ("Licenses for Topsy API" 2012).

For the purposes of this paper, which required tweet content trending over many months, the Topsy API was chosen (using the publicly available key that allowed for three thousand requests per day) , since Twitter does not provide long term retrieval of its tweets.

4.2 Tweet Query Considerations

As discussed in 4.1, the Topsy API was chosen to extract tweets from its archives. All queries were made with a parameterized perl script.

4.2.1 Query Keyword Formation

Tweets can be queried with either bare-keyword or hash tag-keyword encoding. In the case of bare-keyword encoding, a space is represented by a percent symbol (%) followed by 20: %20; for hash tag (#) keywords, a percent symbol followed by 23: %23.

4.2.2 Query String Formation

Topsy's API format allows users to download tweets in the JSON hierarchical format, utilizing parameters for query keywords, requests-per-page, a page counter, and other

more specialized parameters to include or set JSON entities, arrays and hashes for each tweet ("Topy's Otter API ", 2012, June 10). An example of the script call utilizing Topsy's API follows:

otter6.pl <begin_date_offset > < extract_file > <query_key> <timerange> <end_date>
Ex. otter6.pl 1 C:\demos\toyota_tweets.txt Toyota 21600 1

Effectively, this would go back one day from today, use the query keyword *Toyota* and inspect every time range of 21600 seconds (every six hours) within that day for tweets matching that keyword. Since the end date increment is the same as the begin date, it would only do this for one day, but this could be expanded to go back indefinitely.

Obviously, going back too far (more than a month) for a single script call would require additional monitoring because eventually a signal drop-off will occur, and it is also very likely that within the period of one day, the query limit of 3000 queries will be exceeded, and the script will halt. In short, for the purpose of this paper, tweets were extracted using a one day window (begin and end date are the same number), and multiple days were handled in batches, where the begin and end date, as noted above, were incremented by one. This method was utilized to return approximately one and a half years of Toyota tweets, going back from November 2010 to April 2012, for approximately 900,000 Toyota tweets.

4.3 Tweet Loading

Tweet loading was performed by concatenating all of the files created by the queries and then loading them as a single flat file into the mysql database.

4.4 Tweet Uniqueness

It is convenient that each tweet's ID (originally assigned by Twitter) is a fixed number. Based on that attribute alone, its uniqueness remains fixed in the database.

5. CREATING A TERM DOCUMENT MATRIX

A large set of tweets can be gathered into a term document matrix in order to enumerate frequencies, and establish patterns. The general method for doing is to iterate through the entire set of tweets, removing punctuation where necessary to identify the count of individual terms in a sorted order. Filtering through a general stop-word list is necessary so that repetition of common articles, pronouns and other parts of speech that are not relevant will eventually leave a list of sorted relevant terms. Even though the program automates the counting and sorting of relevant terms, it is important to understand the topic domain in order to have the best feel for what types of terms to eliminate (by modifying the script). At the end of running the term document matrix script, the output contains rows that represent individual tweets, with the columns representing the top terms in sorted order. The script that ran on the tweets gathered (between 10/2010 and 04/2012) for this paper yields the following terms and their related frequencies.

Table 5-1

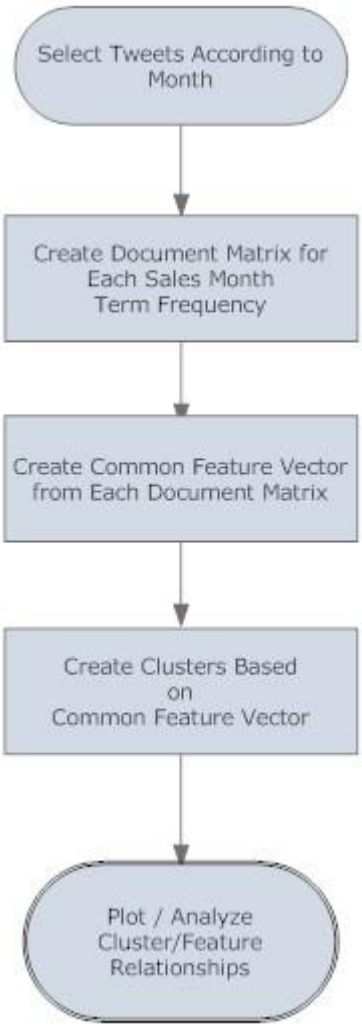
term	count
toyota	897698
new	100440
prius	76259
camry	70008
replacement	61604
2012	54483
side	52960
i	45928
rt	38783
corolla	36821

In the above, the extracted terms reflect the top ten highest frequencies of terms in the entire tweet set. Essentially, this gives a very clear, initial indicator about the relative importance of certain Toyota models and their related terms.

6. CLUSTER ANALYSIS

Cluster analysis of tweets can offer important insights. While sentiment analysis is supervised (utilizing training data through observation and categorization), cluster analysis offers the powerful advantage of data partitioning by assigning sets of data, and in this case, monthly sales data corresponding to tweets, to unique data points within a cluster. The general approach for performing cluster analysis follows.

Figure 6-1



6.1 Clustering Method

As specified in the flowchart, cluster analysis will involve separating the data into document sets that correspond to given Toyota sales months. Using these data sets, the method continues by creating a term document matrix for the highest frequency terms. These highest frequency terms can later be utilized for creating a common set of features against which tweets through the entire sales period can be gathered to form clusters.

6.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a data reduction technique whereby a large set of variables is reduced to a smaller set to determine key patterns and relationships. A thorough and mathematical explanation can be found in many references, but a basic explanation of PCA is that it "reduces variables in a way that as much variability is retained" (Bilosoly, 2008, p.205). A more visual analogy would include measuring the linear motion of a spring from three different angles, and through a linear reduction technique, remove the redundancies from the three measurements, essentially noise, while capturing the variance (Schlens, 2006).

The statistical package *R* contains a function that automates this process by extracting the term document matrix, and then performing the linear reduction, resulting in the principal components (PC prefix) for the various terms. This was done for the eighteen unique document term matrices, utilizing the common term vector for that particular month.

```
ToyotaTerms=read.csv("C:\\Demos\\LoadingScripts\\Cluster\\Gold\\2011-10-01.csv")
ToyotaPrinComp=prcomp(ToyotaTerms, ….)
ToyotaPrinComp->Principal Components Table
write.csv(file="C:\\Demos\\...PrinComp_2010-11-01.txt", ToyotaPrinComp$rotation)
CumProp=summary(ToyotaPrinComp)
```

Repeating the above for all sales months produced eighteen different sets of principal components, one table for each month. An additional *R* calculation performed a *summary* of the Principle components for that month. The *summary* is derived from the Eigen values (contained in the *ToyotaPrincomp$sdev* member variable) , which measure

the amount of variation for each PC, and will be the larger for the first PC and smaller for subsequent PC's (Fernandez, n.d.). Table 6-1 below depicts what the coverage is for each PC by month, for the first four principal components (the data captured from the *summary* function). In many instances, the bulk of the variation is explained in PC1 or PC1 and PC2. The Principal Components or *Eigenvectors* contain the weights provided by the *ToyotaPrinComp$rotation* variable for the various terms A min/max breakdown was provided for each set of principle component weights in Table 6-2 and Table 6-3. These weights indicate what are potential terms of interest for clustering; for example, the Prius *c* has a maximum value for several months and would therefore be a good consideration for using in a cluster. This is not to say that all min/max values were used for clustering, in some instances, such as the term *shareathon*, it was evident by observing the term document matrix - that term had a high count for one month (12/11/2011), and almost non-existent for other months, and that it would be good candidate for clustering, even though it was not a min/max value.

Table 6-1

Month	PC1 Coverage	PC2 Coverage	PC3 Coverage	PC4 Coverage
11/1/2010	0.90	0.98	0.99	1.00
12/1/2010	0.84	0.91	0.95	0.98
1/1/2011	0.75	0.95	0.98	0.99
2/1/2011	0.91	0.97	0.99	0.99
3/1/2011	0.93	0.96	0.99	0.99
4/1/2011	0.89	0.97	0.99	1.00
5/1/2011	0.79	0.92	0.96	0.98
6/1/2011	0.44	0.74	0.92	0.97
7/1/2011	0.66	0.92	0.96	0.97
8/1/2011	0.54	0.84	0.94	0.98
9/1/2011	0.52	0.84	0.94	0.97
10/1/2011	0.55	0.83	0.97	0.99
11/1/2011	0.70	0.90	0.95	0.97
12/1/2011	0.98	0.99	1.00	1.00
1/1/2012	0.55	0.83	0.92	0.97
2/1/2012	0.49	0.86	0.97	0.98
3/1/2012	0.54	0.80	0.92	0.97
4/1/2012	0.50	0.75	0.88	0.96

Table 6-2

Term	Month	PC1 (Min Val)
recal	11/1/2010	-0.019
plugin	12/1/2010	-0.033
c	1/1/2011	-0.018
rt	2/1/2011	-1.000
product	3/1/2011	-0.008
rt	4/1/2011	-0.999
plugin	5/1/2011	-0.033
new	6/1/2011	-0.903
plugin	7/1/2011	-0.004
mani	8/1/2011	-0.011
peopl	9/1/2011	0.008
smartphon	10/1/2011	-0.004
motor	11/1/2011	0.003
v	12/1/2011	0.015
wagon	1/1/2012	0.005
first	2/1/2012	0.029
month	3/1/2012	0.016
year	4/1/2012	0.013

Table 6-3

Term	Month	PC (Max Value)
rt	11/1/2010	1.000
rt	12/1/2010	0.898
rt	1/1/2011	0.898
top	2/1/2011	0.014
rt	3/1/2011	1.000
plugin	4/1/2011	0.045
rt	5/1/2011	0.898
sale	6/1/2011	0.119
sale	7/1/2011	0.578
new	8/1/2011	0.939
new	9/1/2011	0.883
c	10/1/2011	0.770
c	11/1/2011	0.840
rt	12/1/2011	0.555
c	1/1/2012	0.762
new	2/1/2012	0.641
c	3/1/2012	0.754
c	4/1/2012	0.746

6.1.2 Clustering Algorithm

The clustering method for this section utilizes the K-means algorithm. The follow approach summarizes this method (Tan 2006).

Select K points as initial centroids
Repeat
Form K clusters by assigning each point to its closes centroid
Reassign each cluster centroid
Until Centroids do not change

The *R* statistics package contains a *k-means* function (Bilosoly, 2006) that will accept input data and assign it to clusters according to the algorithm above. It will also allow retrieval, by label, the individual data points within each cluster, and plotting of the data. This function will be utilized to render the clusters and plot the individual Toyota sales month data points.

6.1.3 Toyota Sales Data

Complete Toyota monthly sales data (“Toyota Retail Sales”, 2010, Nov.-Dec.; “Toyota U.S. Sales Summary”, 2011; “Toyota U.S. Sales Summary”, 2012, Jan.-Apr.) was recorded for all models, including the Toyota Prius, which will be the focus of this cluster analysis. Toyota sales data was collected in the following table (by unit) for the Prius.

Table 6-4

Year/Mo	MonthNum	Prius
2010/11	1	10224
2010/12	2	15639
2011/01	3	10635
2011/02	4	13539
2011/03	5	18605
2011/04	6	12477
2011/05	7	6924
2011/06	8	4340
2011/07	9	7907
2011/08	10	9491
2011/09	11	9325
2011/10	12	11008
2011/11	13	15208
2011/12	14	17004
2012/01	15	11555
2012/02	16	20593
2012/03	17	28711

6.1.4 Toyota Sales Clustering Results

The clustering method described at the beginning of this section was utilized to explore several different aspects of the Toyota sales data: comparison of Prius model *c* and *v*; the

Shareathon program and recall programs; how the Shareathon program foreshadowed sales of new vehicles; the impact of the March 2011 earthquake and tsunami in Japan.

6.1.4.1 Comparison of Prius c and v Models

From tweet analysis, the most popular Prius models are the Prius c and the Prius v, the former is the compact economical version of the Prius and the latter is the mini-van version. Throughout the tweet analysis, for all sales months, these two models where consistently the top two models tweeted about. The following is the k-means clustering plotted for the tweet-rate of both models (through entire sales period 11/2010 through 04/2012).

Figure 6-2

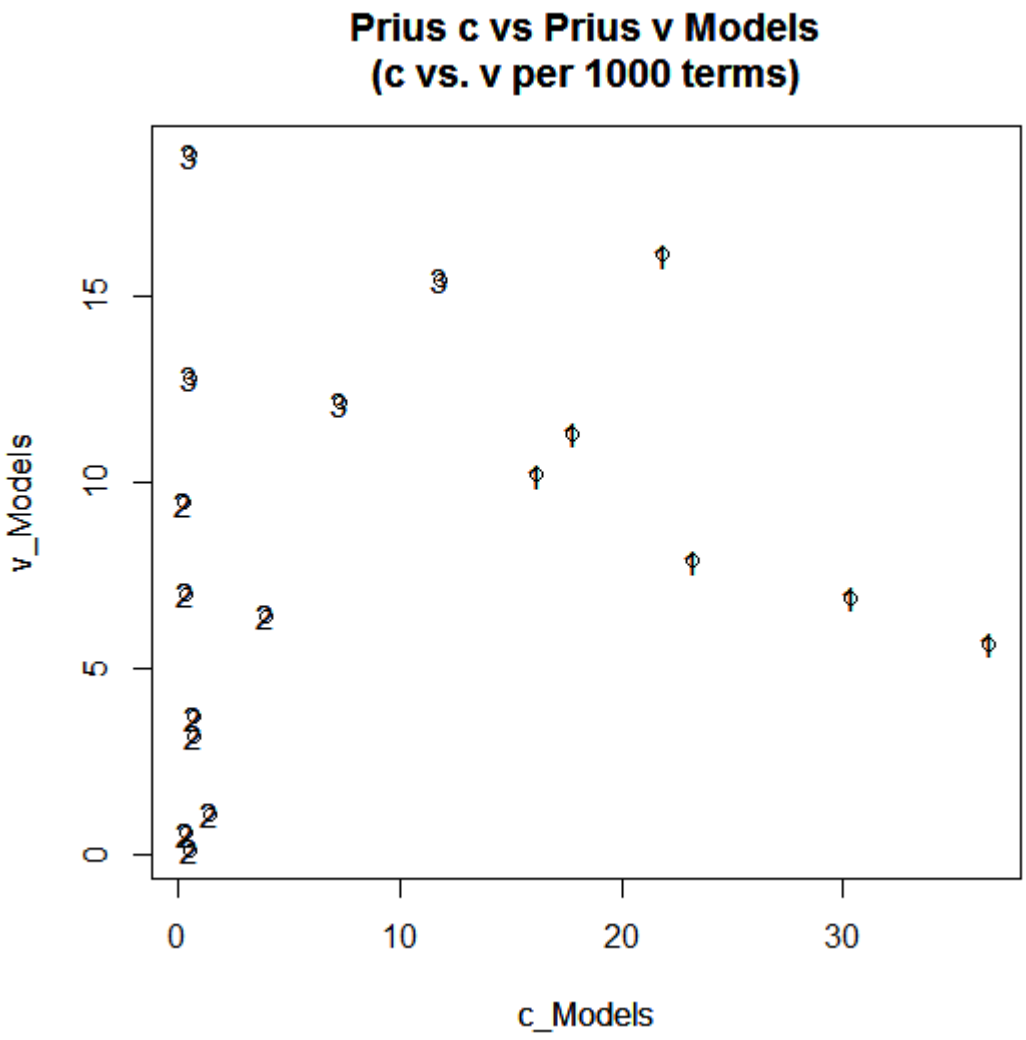

Prius c vs Prius v Models
(c vs. v per 1000 terms)
v_Models
c_Models
0
5
10
15
0
10
20
30

Table 6-5

Cluster	Sales Month
1	**Prius 10/2011 Tweets (11008 Units)**
1	**Prius 11/2011 Tweets (15208 Units)**
1	**Prius 01/2012 Tweets (11555 Units)**
1	**Prius 02/2012 Tweets (20593 Units)**
1	**Prius 03/2012 Tweets (28711 Units)**
1	**Prius 04/2012 Tweets (25168 Units)**
2	Prius 11/2010 Tweets (10224 Units)
2	Prius 12/2010 Tweets (15639 Units)
2	Prius 02/2011 Tweets (13539 Units)
2	Prius 03/2011 Tweets (18605 Units)
2	Prius 04/2011 Tweets (12477 Units)
2	Prius 07/2011 Tweets (7907 Units)
2	Prius 08/2011 Tweets (9491 Units)
2	Prius 12/2011 Tweets (17004 Units)
3	Prius 01/2011 Tweets (10635 Units)
3	Prius 05/2011 Tweets (6924 Units)
3	Prius 06/2011 Tweets (4340 Units)
3	Prius 09/2011 Tweets (9325 Units)

As noted in the later four values in Cluster 1, the Prius *c* Model made significant gains in terms of twitter activity during the first and second quarter of 2012, so its popularity may have contributed to a significant increase in unit sales during that period.

6.1.4.2 Cluster Evaluation of Market Programs vs. Prius Recalls

In addition to the sales of the most popular version of Prius, another important analysis involves comparing Toyota's customer-based marketing program *shareathon* against the Prius recalls over the 11/2010 to 04/2012 sales periods. In order to do this, the actual terms *shareathon* and *recal* were plotted against each other. Note that the latter term *recal* is actually the stemmed, normalized term for the set of terms [*recalls, recalled, recalling]* and so on.

Figure 6-3

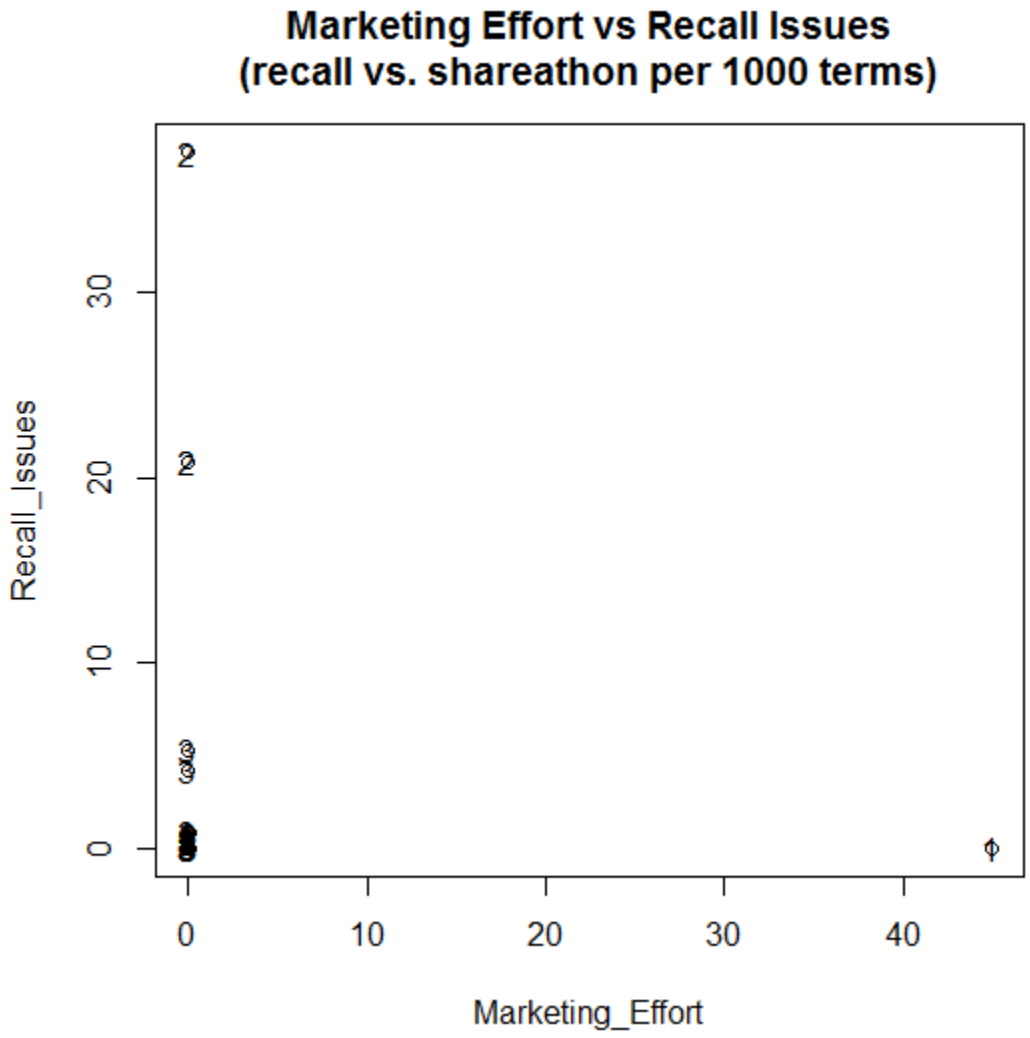

Marketing Effort vs Recall Issues
(recall vs. shareathon per 1000 terms)
Recall_Issues
0
10
20
30
0
10
20
30
40
Marketing_Effort

Table 6-6

Cluster	Sales Month
1	**Prius 12/2011 Tweets (17004 Units)**
2	**Prius 11/2010 Tweets (10224 Units)**
2	**Prius 06/2011 Tweets (4340 Units)**
3	Prius 12/2010 Tweets (15639 Units)
3	Prius 01/2011 Tweets (10635 Units)
3	Prius 02/2011 Tweets (13539 Units)
3	Prius 03/2011 Tweets (18605 Units)
3	Prius 04/2011 Tweets (12477 Units)
3	Prius 05/2011 Tweets (6924 Units)
3	Prius 07/2011 Tweets (7907 Units)
3	Prius 08/2011 Tweets (9491 Units)
3	Prius 09/2011 Tweets (9325 Units)
3	Prius 10/2011 Tweets (11008 Units)
3	Prius 11/2011 Tweets (15208 Units)
3	Prius 01/2012 Tweets (11555 Units)
3	Prius 02/2012 Tweets (20593 Units)
3	Prius 03/2012 Tweets (28711 Units)
3	Prius 04/2012 Tweets (25168 Units)

In terms of tweets, Clusters 1 and 2 represent the bulk of the recall activity and marketing activity for all tweets collected, and one data point from each are represented at the opposing corners of the graph. Of special note is the data point for the 06/2011 Sales month. Digging a bit deeper into the tweet-set for this month, many of the recall tweets were made in reference to a steering issue.

Toyota announces voluntary recall of around 52,000 Prius vehicles to address steering issues

Along the recall lines, the 11/2010 sales month from Cluster 2 contained recall tweets that involve issues with pumps and cooling systems.

Toyota Recalls 650000 Priuses For Faulty Pumps /Toyota To Recall Prius For Cooling Problem

It's interesting to note that 06/2011, the month with the greatest instances of recall tweets, is also the month with the lowest unit sales. Although the abundance of recall tweets in this month appears to account for the low unit sales, further analysis of other issues within the entire 11/2010 through 04/2012 sales period needs to be conducted.

Finally, the *shareathon* marketing effort, exclusive to 12/2011, appears to drown out any deleterious effect of recall notices by sending a strong signal to retweet about buying a new Prius.

> This tweet s worth $500 + $50 per RT if I get a new Toyota during Toyotathon. RT for a chance at a Prius. Shareathon

In a sense, the marketing program provides a positive guard-band against negative sentiment generated by recall tweets that were massively re-tweeted as well. The marked year-to-year increase in sales during 01/2012-04/2012, would indicate that the customer-based incentive to market the Prius worked.

6.1.4.3 Using the "New" Term to Foreshadow Next Year's Sales

Plotting the term *new* against Prius *c* tweets demonstrates yet another interesting relationship. Again, tweets generated from the December 2011 Shareathon program herald the upcoming year's models, including Prius c. As mentioned in the previous section, the simple fact that year-year sales increased markedly for the Prius in the 01/2012 to 04/2012 sales months leave at least a general indicator that this marketing approach works.

Figure 6-4

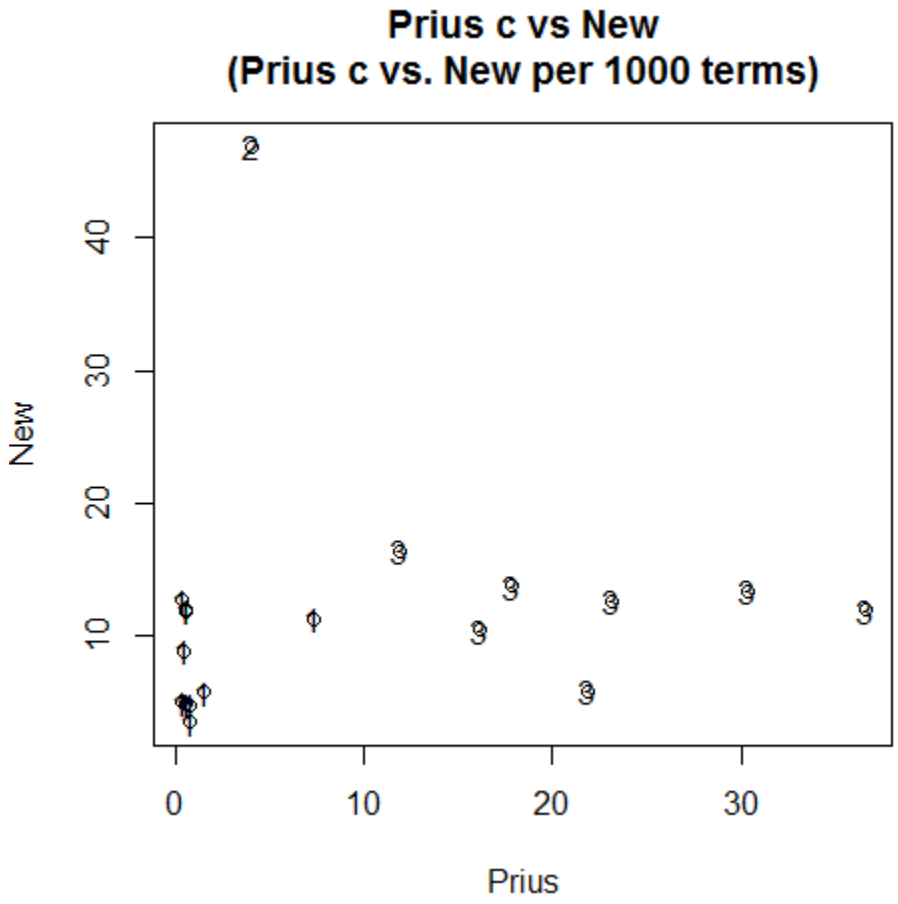

Prius c vs New
(Prius c vs. New per 1000 terms)
New
Prius
0
10
20
30
10
20
30
40

Table 6-7

Cluster	Sales Month
1	Prius 11/2010 Tweets (10224 Units)
1	Prius 12/2010 Tweets (15639 Units)
1	Prius 02/2011 Tweets (13539 Units)
1	Prius 03/2011 Tweets (18605 Units)
1	Prius 04/2011 Tweets (12477 Units)
1	Prius 06/2011 Tweets (4340 Units)
1	Prius 07/2011 Tweets (7907 Units)
1	Prius 08/2011 Tweets (9491 Units)
1	Prius 09/2011 Tweets (9325 Units)
1	Prius 05/2011 Tweets (6924 Units)
2	**Prius 12/2011 Tweets (17004 Units)**
3	**Prius 01/2011 Tweets (10635 Units)**
3	**Prius 10/2011 Tweets (11008 Units)**
3	**Prius 11/2011 Tweets (15208 Units)**
3	**Prius 01/2012 Tweets (11555 Units)**
3	**Prius 02/2012 Tweets (20593 Units)**
3	**Prius 03/2012 Tweets (28711 Units)**
3	**Prius 04/2012 Tweets (25168 Units)**

6.1.4.4 The Impact of the 2011 Earthquake

In March 2011, Japan experienced an earthquake and tsunami that had a devastating impact on a large segment of the population, and substantially crippled its manufacturing and business infrastructure. As would be expected, it created delays throughout all industries that relied on parts for end-use manufacturing. For Toyota, this impacted the manufacture and shipment of the *Prius*. The twitter datasets for the 03/2011 sales months documented this effect.

Figure 6-5

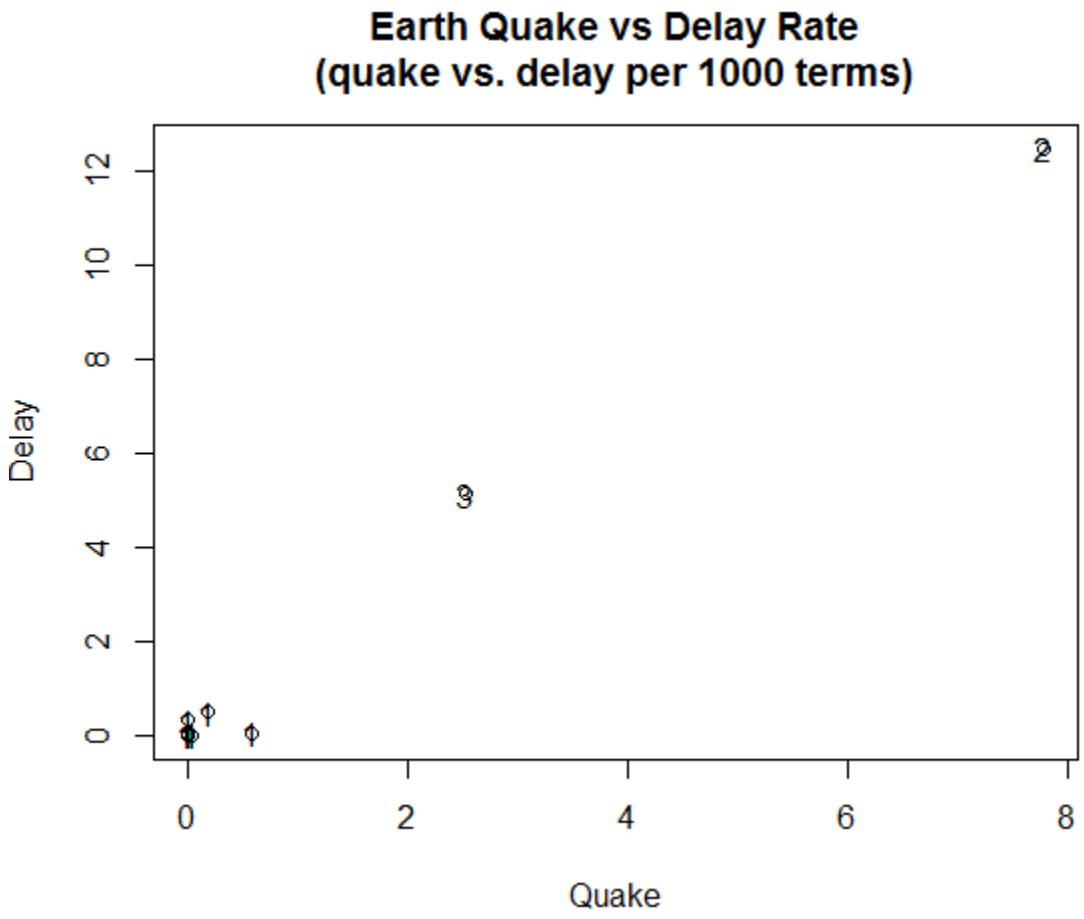

Earth Quake vs Delay Rate
(quake vs. delay per 1000 terms)
Delay
0
2
4
6
8
10
12
Quake
0
2
4
6
8

Table 6-8

Cluster	Sales Month
1	Prius 11/2010 Tweets (10224 Units)
1	Prius 01/2011 Tweets (10635 Units)
1	Prius 02/2011 Tweets (13539 Units)
1	Prius 03/2011 Tweets (18605 Units)
1	Prius 04/2011 Tweets (12477 Units)
1	Prius 06/2011 Tweets (4340 Units)
1	Prius 07/2011 Tweets (7907 Units)
1	Prius 08/2011 Tweets (9491 Units)
1	Prius 09/2011 Tweets (9325 Units)
1	Prius 10/2011 Tweets (11008 Units)
1	Prius 11/2011 Tweets (15208 Units)
1	Prius 12/2011 Tweets (17004 Units)
1	Prius 01/2012 Tweets (11555 Units)
1	Prius 02/2012 Tweets (20593 Units)
1	Prius 03/2012 Tweets (28711 Units)
1	Prius 04/2012 Tweets (25168 Units)
2	**Prius 03/2011 Tweets (18605 Units)**
3	**Prius 05/2011 Tweets (6924 Units)**

In addition to the two months highlighted, it is important to note that 06/2011, which had a disproportionately high number of recall tweets, was also the lowest in unit sales for the entire period, but that could also be accounted for by the effects of the earthquake.

7. CONCLUSION

This paper analyzed social networking signals (tweets) for a given car company, Toyota, in terms of their impact on sales. The approach and method of this paper addresses questions posed every day by top level management representing a given company's finance and marketing department:

> *What is the composite sales picture of our products, including all of the drawbacks, marketing program successes and unexpected events? What indicators allow us to track, compile and evaluate these events?*

The approach followed in this paper answers these questions, and since it was, to a large extent, programmatically developed, it could be productized to fit many sales environments where social media events and data are measured through time. The following sections illustrate more detail of why this is so.

7.1 API Query Method for Extracting Twitter Data

As described in Sec. 4, in order to gather a comprehensive set of a data, an API based query script employing the term *toyota* (and nothing else) extracted all related tweets from 11/2010 until 04/2012, approximately 900,000 tweets total. This approach provided definite advantages, one of which was to include all possible term combinations with *toyota* for that timeframe. Earlier attempts did include more specific query combinations, such as *toyota & rav4* or *toyota & prius* or *toyota & lexus*. The problem

with this approach (ultimately abandoned) is twofold: 1) possible overlap of datasets from different queries 2) partial or total elimination of datasets that contained relevant or compelling information.

In short, starting with a single query term over a given timeline returned a comprehensive, unsupervised picture of all events associate with *toyota*. One added benefit of this approach is consistency: the same approach could be utilized for any entity, corporation, or organization, in order to productize this entire method.

7.2 First Pass Filtering

The next step in analyzing the data was to create a term document matrix by counting the occurrence of terms in the entire tweet set (10/2010 through 04/2012). Terms such as *prius, new, and 2012* were among the highest in frequency count. This gave a clear indication that *prius* was a leading term throughout the entire time period, and additional querying could be done to analyze any events or issues associated over that time span. This is where the unsupervised approach would need to become mixed with the supervised approach of term recognition: for example, as these terms are collected, it is important to know that *prius* is in fact a car model (of many models) whose sales picture is worthy of further research.

7.3 Identifying Key Relationships through Term Clustering

Since tweets could be separated into monthly buckets by their tweet post date, distinct term documents could be created for each month (i.e., the book of "December Prius Tweets", etc.). Utilizing this structure, a term vector could be composed for each month,

and a union of all term vectors could be constructed as well. An important note is that the final production of all tweets, prior to clustering, included the application of stop-word lists and stems filters, so the terms collected were normalized and relevant. Because some tweet characterization was already conducted, some term vectors were selected by direct observation of the data. Alternatively, Principal Component Analysis (PCA) was also utilized to select a set of terms around which a cluster could be built. In the latter, a more structured approach is utilized, because the weights of the Principle Component values and their proportional influence could be evaluated to select the terms around which a cluster could be built.

Ultimately, the results yielded useful and compelling information regarding the sales picture for the entire time period. For example, some months, such as 11/2010, which did not have a high tweet count, had nonetheless very useful information regarding recalls for the Prius model, and could easily be grouped with 06/2011, which had a significant recall count as well. Other clusters detailed the effects of the March 2011 earthquake and tsunami, and the possible effects that it had on sales (of note is that June 2011 had the lowest sales, three months after the quake). And the concentrated marketing approach (the Shareathon tweet program) that was launched and completed in 12/2011, was (apparently) enough to counterbalance other tweet signals throughout the year, that included recalls and unforeseen natural disasters.

In a sense, the clusters tell a compelling story of the events to the detriment and success of the Prius over the entire period. The plain fact that the Prius had record setting year to year Prius sales in 2012, is an indication that the marketing approach

appeared to be successful. However, to exclude the other events that could contribute to this success would be naïve at best.

7.4 Improvements on Overall Method

Stated from the outset of this section, the purpose of this paper was to provide an answer to common industry questions that are asked every day: *what events are revealed through social networking that drive or detract from sales?* This paper outlined and implemented a successful approach to answer that question. However, the approach could be vastly improved by reorganizing some of the applied method. For example, a more common sense approach would have clustering and perhaps Principal Component Analysis precede attempts at classification, and then from those results, better, more representative terms could be used to build training sets for classification that would follow.

Furthermore, it is evident that PCA and clustering can be conducted fairly quickly, and a good characterization of terms and relevant relationships could be established through automation. In fact, to productize this approach completely, Principal Component Analysis could be scripted to analyze terms given various parameters for Eigen value thresholds and cumulative representation. This could effectively be done at the back end, and the implementation could include a graphical interface that would provide user controls to select those thresholds upon observing the graphical results of clustering.

Finally, in order to have a more detailed understanding of how the twitter signal affected sales, other variables affecting the demand of buying a car in the Prius class of

vehicle would need to be considered as well, and could be addressed in a regression algorithm by analyzing car sales against tweet sentiment and other factors, such as the price of gas, unemployment, the behavior of the stock market, and so forth. This part would be more difficult to productize in a generic sense, because any given product in a given market may require careful analysis before determining what variables and diagnostics to apply before a valid regression model could be constructed. Nonetheless, a solid understanding of the key social networking terms and their relationships, through the methods described in this paper, provides a good first step in developing such a model.

8. FURTHER DIRECTIONS

This paper has analyzed tweets in terms of targeted marketing and sales. To a great extent, the tweets themselves are mini-pressrooms that trumpet a company's success, announce new innovations and herald cutting edge ideas. On the other hand, tweets can express negative marketing sentiment that can affect the public's perception of a product.

Regardless of the Twitter message tone, marketing firms, data mining professionals and statisticians will always be interested in understanding different ways to analyze tweets, for as long as Twitter is a service. To that end, analysts have the full range of the Twitter database, which can include examining retweets and analyzing the inclusion of encoded urls, which appear in a substantial number of tweets (if not a majority of them), in order to obtain a finer level of meaning. Both of these approaches are briefly described in the following sections.

8.1 Examining Encoded URLS

Throughout this paper, the actual encoded urls within tweets were largely ignored. In fact, the method developed for scoring tweets removes the "http" string altogether, so it will not be scored as a document term, thus improving the overall efficiency of the program while supposedly not losing any meaning. However, imagine that both positive and negative tweets included encoded urls such as *http://t.co/u5w3ccNq* and *http://bit.ly/g8SoOt* . The first tweet is associated with a "positive" tweet because it is associated with the Toyota Shareathon program, and the second is associated with a negative tweet referring to Toyotas response to lawsuits. Both urls were concatenated together and analyzed by submitting their HTML encoded versions to an API provided by a third party URL decoding website, LongURL ("LongURL API 2.0 Documentation," 2012).

find_url.pl encodedurls.txt……

Original URL --------->http://t.co/u5w3ccNq
HTML encoded ------>http%3A%2F%2Ft.co%2Fu5w3ccNq
Call: :
http://api.longurl.org/v2/expand?url=http%3A%2F%2Ft.co%2Fu5w3ccNq&format=json
Result: http://www.toyota.com/shareathon/seed/rcasey2012/

Original URL -------->http://bit.ly/g8SoOt
HTML encoded ------>http%3A%2F%2Fbit.ly%2Fg8SoOt
Call:
http://api.longurl.org/v2/expand?url=http%3A%2F%2Fbit.ly%2Fg8SoOt&format=json
Result: http://www.upi.com/Top_News/US/2011/04/20/Toyota-fights-amending-crash-death-suit/UPI-13561303325942

In both instances, the URL was decoded and shown to be consistent with the content of the tweet. This approach, however, could be extended to examining encoded urls where

the tweet is otherwise not readily discernible as negative or positive. From there, either by examining the text within the URL, or by perhaps comparing the real URL to one that is indexed and rated in list, the overall tone of the tweet could be scored with more confidence. This approach does add more overhead and can create additional performance issues, but is definitely an additional approach for determining tweet sentiment.

8.2 Examining Retweets

Although this paper did not analyze the retweet count for a given URL, the information is available, and can be extracted by using standard Twitter API tools. The one drawback is that for the regular (non-OAuth) API, a program may only make up to 150 calls per hour, and because each call is unique (unlike a single query that can download many datasets), it is timely and costly to make the over eight hundred thousand calls it would take cover all of the tweets utilized for this paper. An example of this would be a json-based query program ("Tweet Statuses API," 2012) that would return the following content for a Toyota tweet .

retweeted => 0
source => web
favorited => 0
coordinates =>
place =>
retweet_count => 10
truncated => 0
created_at => Sat Jul 23 05:35:19 +0000 2011
in_reply_to_status_id_str =>
contributors =>
text => Race Day tomorrow! Come out to Toyota Speedway at Irwindale and help us raise money for Kids Against Cancer!!
user => HASH(0x3a19b40)
in_reply_to_user_id =>
id => 94641722864304128

In short, Twitter provides rich additional detail for tweets, provided that there is enough time and resources to gather this additional information.

9. SYSTEM DETAILS

9.1 Perl Interpreter

v5.10.0 built for MSWin32-x86-multi-thread
Binary build 1002 [283697] provided by ActiveState http://www.ActiveState.com
Built Jan 10 2008 11:00:53
GNU General Public License, which may be found in the Perl 5 source kit.

9.2 Python Interpreter

Python 2.7.2 (default, Jun 12 2011, 14:24:46) [MSC v.1500 64 bit (AMD64)] on win32

9.3 MySQL / APACHE / PHP

XAMPP Package
+ Apache 2.2.21
+ MySQL 5.5.16 (Community Server)
+ PHP 5.3.8 (VC9 X86 32bit thread safe) + PEAR

9.4 R Open Source Statistical Package

R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

9.5 System Specs

HP Paviliion dm4 Notebook PC
Processor:Intel(R)Core(TM)i5CPU
M450 @240GHz, @2.40GHz
Operation System : Windows 7 (64-bit)

10. REFERENCES

[1] Bilosoly, R. (2008). Multivariate Techniques with Text. *Practical Text Mining with Perl* (pp. 206-211). Hoboken, NJ : J. Wiley & Sons, Inc.

[2] Bilosoly, R. (2008). Clustering. *Practical Text Mining with Perl* (pp. 220-233). Hoboken, NJ : J. Wiley & Sons, Inc.

[3] Fernandez, George (n.d.). Principal Component Analysis, (pp.1-4).. Retreived from http://www.ag.unr.edu/saito/classes/ers701/pca2.pdf

[4] Geng, Y. (2010, October 29-31). The Research of Data Mining Sales Forecast, *International Conference on Multimedia Technology (ICMT). 1-3.*

[5] Holmes, B. (2012, December 11). 5 Ways Social Media Will Change the Way You Work in 2013. *Forbes.* Retrieved from http://www.forbes.com/sites/ciocentral/2012/12/11/5-ways-social-media-will-change-the-way-you-work-in-2013/

[6] Institute of Business Forecasting and Planning. (2012). Retreived from http://ibf.org.

[7] [Licenses for Topsy API] . (2012). Retrieved from http://manage.topsy.com/app

[8] LongURL API 2.0 Version Documentation .(2012). Retrieved from http://longurl.org/api

[9] Mining media to find vehicle defects. (2012 Fall). *Pamplin : College of Business Magazine.* Retrieved from http://www.magazine.pamplin.vt.edu/fall12/vehicledefects.html

[10] Schlens, Jonathon (2005). A Tutorial on Principal Component Analysis, (pp.1-4). Retreived from http://www.snl.salk.edu/~shlens/pca.pdf

[11] Shahid, S. & Manarvi, I. (2009, July 6-9). A Methodology of Predicting Automotive Sales Trends through Data Mining, *Conference on Computers & Industrial Engineering, 2009. CIE 2009. International,. 1-3.*

[12] Tan, Pan-Ning, Steinbach, M. & Kumar, Vipal (2006). Cluster Analysis: Basic Concepts and Algorithms. In Harutunian,K. & Goldstein, M. (Eds.), *Introduction to Data Mining,* , p. 497. Boston, MA: Pearson Education, Inc.

[13] The Streaming APIs. (2012). Retrieved from https://dev.twitter.com/docs/streaming-apis.

[14] Topy's Otter API . (2012, June 10). Retrieved from http://code.google.com/p/otterapi/wiki/Resources#The_Otter_URI

[15] Toyota Announces Second Annual Shareathon Program.(2011, December 11). Retrieved from http://pressroom.toyota.com/releases/toyota+announces+second+annual+shareathon+program.htm

[16] TOYOTA RETAIL SALES. (2010,November). Retrieved from http://pressroom.toyota.com/images/document/Monthly_Sales_PR_Chart.pdf

[17] TOYOTA RETAIL SALES. (2010, December).Retrieved from http://pressroom.toyota.com/images/document/December_10_Sales_Chart.pdf

[18] TOYOTA RETAIL SALES. (2011, January).Retrieved from http://pressroom.toyota.com/releases/january+2011+sales+chart.download

[19] TOYOTA RETAIL SALES. (2011, February).Retrieved from http://pressroom.toyota.com/releases/february-2011-sales-chart.download

[20] TOYOTA RETAIL SALES. (2011,March). Retrieved from http://pressroom.toyota.com/article_download.cfm?article_id=2955

[21] TOYOTA RETAIL SALES. (April,2011). Retrieved from http://pressroom.toyota.com/releases/april+2011+sales+chart.download

[22] TOYOTA RETAIL SALES.(2011,May). Retrieved from http://pressroom.toyota.com/releases/may+2011+sales+chart.download

[23] TOYOTA RETAIL SALES. (2011, June). Retrieved from http://pressroom.toyota.com/releases/june+2011+sales+chart.download

[24] TOYOTA RETAIL SALES. (2011, July). Retrieved from http://pressroom.toyota.com/releases/july+2011+sales+chart.download

[25] TOYOTA RETAIL SALES**. (**2011,August**).** Retrieved from http://pressroom.toyota.com/releases/august+2011+sales+chart.download

[26**]** TOYOTA RETAIL SALES. (2011,September). Retrieved from http://pressroom.toyota.com/releases/september+2011+sales+chart.download

[27] TOYOTA USA Newsroom.(2012).Retrieved from http://pressroom.toyota.com

[28] TOYOTA U.S. SALES SUMMARY. (2011, October).Retrieved from http://pressroom.toyota.com/releases/october+2011+sales+chart.download

[29] TOYOTA U.S. SALES SUMMARY. (2011,November). Retrieved from http://pressroom.toyota.com/releases/november+2011+sales+chart.download

[30] TOYOTA U.S. SALES SUMMARY (2011, December). Retrieved from http://pressroom.toyota.com/releases/december+2011+sales+chart.download

[31] TOYOTA U.S. SALES SUMMARY. (2012, January). Retrieved from http://pressroom.toyota.com/releases/january+2012+sales+chart.download

[32] TOYOTA U.S. SALES SUMMARY. (2012, February). Retrieved from http://pressroom.toyota.com/releases/february+2012+sales+chart.download

[33] TOYOTA U.S. SALES SUMMARY. (2012, March). Retrieved from http://pressroom.toyota.com/releases/march+2012+sales+chart.download

[34] TOYOTA U.S. SALES SUMMARY. (2012, April). Retrieved from http://pressroom.toyota.com/releases/april+2012+sales+chart.download

[35] Tweet Statuses API. (2012). Retrieved from https://dev.twitter.com/docs/api/1/get/statuses/show/%3Aid

[36] Using the Twitter Search API. (2012). Retrieved from https://dev.twitter.com/docs/using-search

[37] Zhang,Z. & Li, X. (2010, March 11). Collective Knowledge of the Web : source of information of process of Business Intelligence, *Proceedings of the 43rd Hawaii International Conference on System Sciences – 2010, Honolulu Hawaii.* 1-2. doi: 10.1109/HICSS.2010.121.