Social Media Analytics:
Data Mining Applied to Insurance Twitter Posts

Roosevelt C. Mosley Jr., FCAS, MAAA

Abstract

The use of social media has grown significantly in recent years. With the growth in its use, there has also been a substantial growth in the amount of information generated by users of social media. Insurers are making significant investments in social media, but many are not systematically analyzing the valuable information that is resulting from their investments.

This paper discusses the application of correlation, clustering, and association analyses to social media. This is demonstrated by analyzing insurance Twitter posts. The results of these analyses help identify keywords and concepts in the social media data, and can facilitate the application of this information by insurers. As insurers analyze this information and apply the results of the analysis in relevant areas, they will be able to proactively address potential market and customer issues more effectively.

Keywords. Social media, analytics, data mining, text mining, clustering, association analysis

1. INTRODUCTION

Regardless of where you look, you can see an explosion in the use of social media. Online communities have developed that focus on both personal and professional lives. Groups have been formed that focus on every potential area of interest, including food, sports, music, parenting, scrapbooking, and actuarial issues. It is estimated that there are over 900 social media sites on the internet. Some of the more popular platforms are Facebook, Twitter, LinkedIn, Google Plus, and YouTube. To help understand the explosion in the use of social media, consider the following statistics which were compiled at www.Banking2020.com [1] in January 2011 and by Danny Brown on his blog at www.dannybrown.me [2].

- People spend over 500 billion minutes per month on Facebook.
- There are 200 million registered Twitter accounts.
- There are more than 70 million users of LinkedIn worldwide.
- YouTube receives more than 2 billion viewers per day.
- Seventy-seven percent of internet users read blogs.

The majority of the population is using social media in some form or another. Given the substantial increase in the use of social media, there is a significant amount of information that is being generated. As seen in the same sources referenced above, the volume of this content is staggering:

- More than 30 billion pieces of content are shared each month on Facebook.
- Every minute, 24 hours of video is uploaded to YouTube.
- As of December 2010, the average number of tweets sent per day was 110 million.
- There are currently 133 million blogs listed on leading blog directory Technorati.

So not only are people joining and accessing social media sites, but they are also spending time engaging in social media and creating a significant amount of content. As a result of this time spent on social media and the information being generated, businesses have taken notice and are attempting to leverage the power of social media to help them succeed. According to wealthinvest.com [3],

- 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.
- Eighty-eight percent of companies use LinkedIn as a recruitment tool.
- Corporate blogging accounts for 14% of blogs.

The commitment that businesses are making to increase their presence in social media is also being shown in the resources they are committing to this effort. According to eMarketer, U.S. marketers will spend over $3 billion to advertise on social media sites in 2011, which is a 55% increase over what was spent in 2010, and 11% of what they spend on online advertising overall. Also, according to Banking2020.com, 50% of Chief Marketing Officers at Fortune 1000 companies say they have launched a corporate blog because it is a cost of doing business today. So not only is the corporate investment being evidenced by dollars spent but also in the time it takes to create and maintain social media efforts.

Insurance companies have joined in this effort by businesses to use social media. The Customer Respect Group (CRG) produces a monthly newsletter entitled "Social Eyes: The Insurer's View of Social Media." [4] This newsletter focuses on trends and news related to the use of social media by insurance companies. As part of the newsletter, CRG tracks the use of insurer social media sites.

One of the categories that is tracked is the use of Facebook. In the June, 2011 issue of "Social Eyes," CRG tracks 36 insurance company corporate Facebook pages that have a collective total of over three million fans. CRG also tracks other insurance company-related Facebook pages, such as advertising personalities or pages that are targeted toward a specific demographic. These ten Facebook pages have over 5.7 million fans, with the largest being the Facebook page of Flo, the Progressive Girl, who has just over 3 million fans. There is also a section in the newsletter that tracks the Twitter followers and activity of 30 corporate insurance company Twitter identities.

CRG describes the different ways that social media is being used by insurance companies. Insurance companies are using social media for broad purposes, such as communicating general content and promotional advertising, but some are also using the platform to contact and communicate with customers directly. In addition, insurance companies are using social media as a platform for promoting and raising money for charities, i.e., donating to a particular cause for every new fan or for every new Twitter follower.

The use of social media sites has grown significantly, and this fact is being recognized by businesses, including insurance companies. In response, insurance companies and other businesses are investing significant time and resources into establishing and maintaining a social media presence. All of this is feeding into the exponential increase in the amount of data and information that is being generated. This then raises a significant question. What are companies in general, and insurance companies specifically, doing with all of this information? Every Facebook post, every Tweet, every blog entry, every connection with social media generates a new data point, a new bit of information that may be of value to insurance companies. This information might help an insurance company service a policyholder better, connect with a potential new customer, identify a need or concern in the marketplace, or uncover a competitive issue they may be facing. Social media platforms provide opportunities for consumers to share their thoughts with a broader audience, and in understanding how customers are feeling or what they are facing, insurers can better interact with consumers.

How can insurers access and begin to make use of this information? Based on the statistics above related to the amount of social media content that is available, the task can be overwhelming. One approach is for a company to hire a team of people to monitor social media sites for content that might be valuable to a company. This would require the team to read through content, identify information that the company may be able to use, and then to identify the proper channels through which to route that information. While this can and is being done, the obvious challenge is that as the amount of social media content continues to increase, more human resources will be required to analyze it. This means either increasing the budget for social media efforts, or simply living with the fact that you cannot analyze everything, thus potentially missing valuable information.

The purpose of this paper is to describe, through the use of a specific example, how data mining and analytics can be applied to social media. Data analytics provides insurance companies with a systematic, organized, and powerful way to analyze social media information and extract the valuable information in it without needing to read through every piece of content. Using the power of analytics, key areas of importance can be identified, and these areas can then be investigated further. This can optimize the time spent by focusing the analysis on those areas that are of the greatest potential benefit to the company.

1.1 Research Context

The context of this research will focus on data and text analytics. Since much of the data from the social media sites will be text-based data, the process of preparing and analyzing the data will focus on principles of preparing text data for analysis. The author was unable to find anything in CAS literature that focuses specifically on the analysis of social media. However, a good discussion of the principles of text mining in CAS literature is in a paper written by Louise Francis entitled "Taming Text: An Introduction to Text Mining." [5] Building on these concepts, there are some unique considerations when analyzing text data from social media sites which will be discussed in this paper.

1.2 Objective

The purpose of this paper is to describe, through the use of a specific example, how data mining and text analytics can be applied to social media to identify key themes in the data. Specifically, this paper will describe the analysis of Twitter posts related to the keyword Allstate. Allstate was chosen purely based on the public availability of historical Twitter data. While this example helps to make some of the points and concepts clearer, the purpose of this paper is not to provide a detailed analysis of Twitter activity related to Allstate, but to demonstrate how analytics can specifically be applied to social media information related to a property and casualty insurance company.

1.3 Outline

Section 2 will provide a general background and description of Twitter and will describe the data used in this analysis. Section 3 will provide some general descriptive statistics about the data. Section

4 will discuss the steps that were taken to prepare the data for analysis. Section 5 will describe the analysis of the tweets and also provide some insight into the results of the analysis. Section 6 will outline some of the challenges associated with analyzing social media information. Finally, Section 7 will give applications of these types of analyses for insurance companies.

2. TWITTER BACKGROUND AND DATA DESCRIPTION

Twitter is a social networking site that allows users to send and read short messages of a maximum of 140 characters. Twitter was created in March 2006 and was officially launched in July 2006. The growth of Twitter has been phenomenal, currently having reached over 200 million users and handling over 200 million tweets per day. Users sign up for an account on Twitter, and once they have an account they can begin to "tweet," which is the terminology for sending a message. Users can subscribe to other user's tweets, a process known as "following." These subscribers are known as "followers." By default, tweets that a user sends are visible to everyone; however, users can also choose to send tweets specifically to their followers that will not be visible to the public.

Users on Twitter are identified by a user name, and this user name is proceeded by the "@" symbol. When a user identifies another user in their tweet by their user name, it will be visible to the public, and the user that is referenced will be notified by Twitter that they have been "mentioned." If a user sees a tweet that is interesting and wants to pass the information along, they can "retweet" the post, which is similar to forwarding an email message to a new set of users, in this case their followers. Retweets will generally be identified with an "RT" that is embedded in the message. Lastly, messages can be grouped by topic or type by the use of hashtags (#). A hashtag preceding the topic will allow Twitter users to find tweets related to a particular topic when performing a search.

Twitter also has a location function. If users are tweeting from a mobile device, they can choose to turn on their location, and their latitude and longitude will be captured with the tweet.

Tweets can be related to anything, but much of the content on Twitter is related to several key categories. These categories were outlined in research done by Pear Analytics in 2009 on 2,000 tweets [6]. This study found that tweets were primarily related to six categories:

1. Pointless babble – 40%
2. Conversational – 38%

3. Pass along value – 9%
4. Self-promotion – 6%
5. Spam – 4%
6. News – 4%

While these numbers are related to a study that was done two years ago when Twitter was not as widely used as it is now, the general categorization of tweets likely still holds. As it relates to insurance companies, the areas of interest would be categories 2, 3, and 6, which account for 51% of tweets. Certainly, not all tweets in these categories will be useful to insurers, so the challenge is to determine how to analyze the tweets in such a way that the important information is separated from the information that is not important.

To demonstrate how this can be done, a dataset of insurance company tweets was identified for analysis. Twapperkeeper.com was a web service that tracked and archived Twitter posts based on archives that were set up by users. To track tweets related to a particular topic or user, users could go to this site and establish an archive, and twapperkeeper.com would track and archive those tweets. On July 29, 2010, an archive titled #allstate was established. Tweets for this hashtag were collected from this starting point, with the exception of about a five-week time period (to be explained later). This analysis used tweets through August 12, 2011. [Author's Note: Since the paper was completed, twapperkeeper.com was fully integrated into hootsuite.com. This site allows users to archive social media data based on defined search criteria. However, archives established by other users cannot be accessed unless explicit permission is given by the owner of the archive.]

The data that was captured from twapperkeeper.com includes the following information:

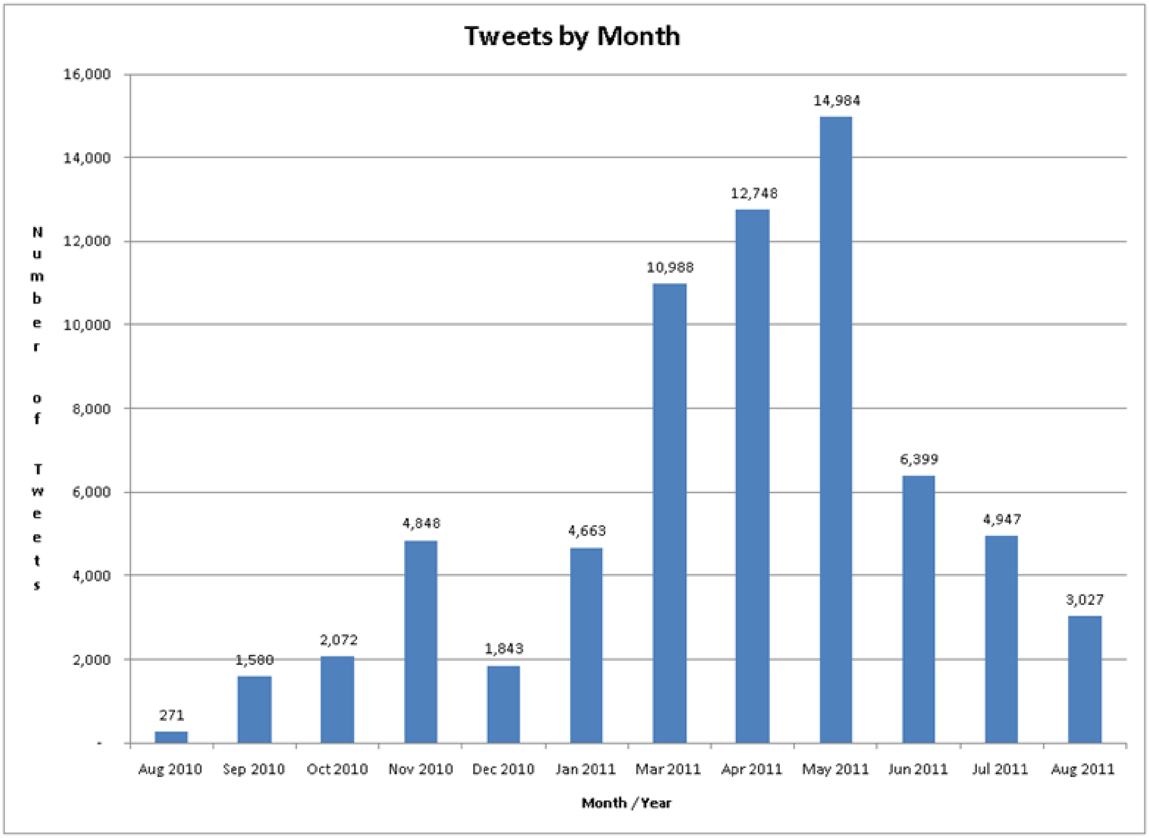
- User: the username that sent the tweet
- Tweet: the content of the tweet
- Timestamp: the date and time the tweet was sent (GMT)
- Tweet ID: Twitter identification number of the tweet
- Geo: latitude and longitude of the user

It must also be remembered that this data was captured based on the use of the hashtag Allstate. Therefore, it will not capture every tweet that uses the word Allstate, but rather those tweets where the user specifically identified Allstate as a keyword. Also, twapperkeeper.com makes no guarantees that they capture all tweets that meet the archive criteria, so there could potentially be tweets with #allstate that were not captured. While this may introduce a bias, the concepts for analyzing the tweets are still valid.

3. GENERAL DESCRIPTIVE STATISTICS

There are a total of 68,370 tweets that were used as part of this analysis. The tweets used began on August 1, 2010 and ended on August 12, 2011. The number of tweets by month is shown below.

Figure 1: Tweets per Month



As can be seen in the figure above, the number of tweets captured per month varied between 1,500 and just under 5,000 through January 2011, at which point the number of tweets increased to 10,000 – 15,000 per month for March through May 2011. June and July settled back to pre-March, 2011 levels. August 2011 only represents 12 days of tweets, so it was not a complete month as of the time this paper was written.

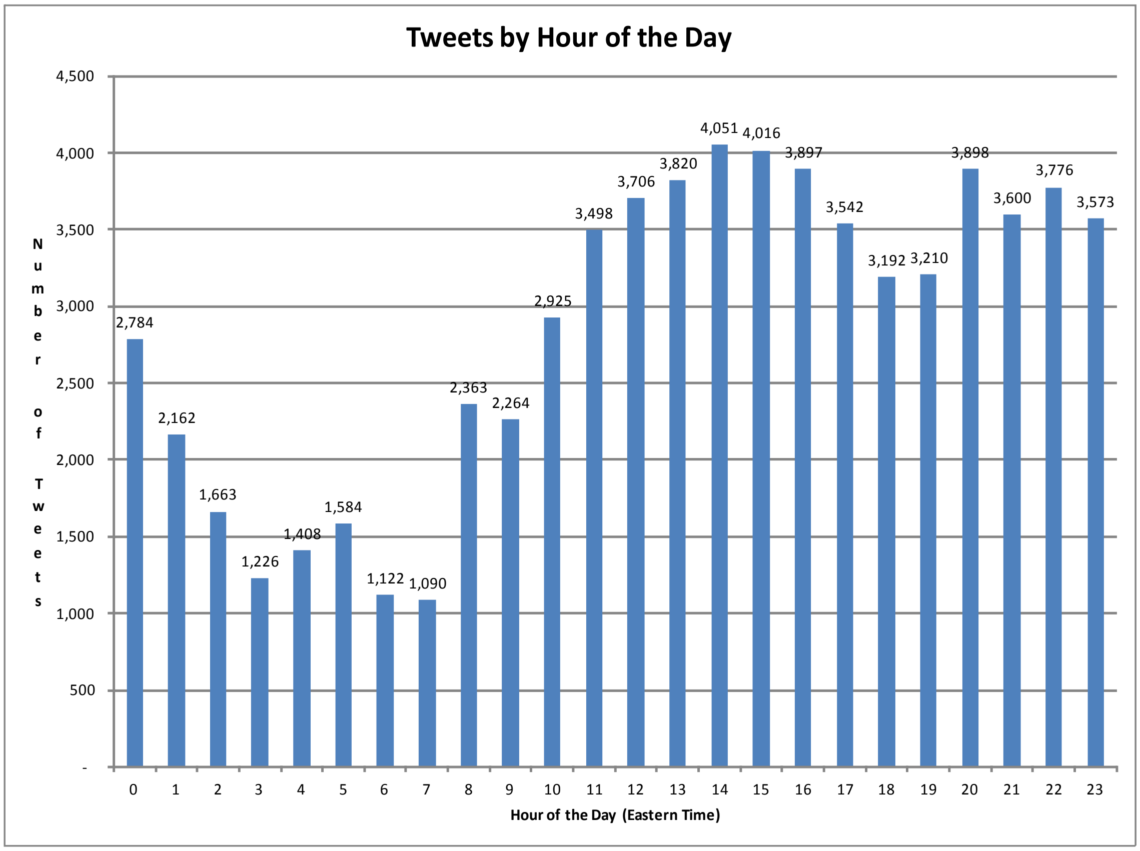
Throughout this paper, as will be described later, we will attempt to uncover concepts being communicated in the data through the use of various data mining techniques. Generally, what the uncovering of these realities does is create more questions which require more investigation to come up with more complete answers. There are several questions which are apparent from Figure 1. First, there seemed to be a significant increase in the number of tweets per month from March to May 2011, followed by a decrease. In addition, as mentioned earlier, there is about a five-week period during which no tweets were captured. We will address the issue related to the increase in tweets later in this paper. Regarding the missing tweets, the author contacted twapperkeeper.com to inquire about the missing data. Early in 2011, there were some issues with the archiving servers, and as a result some of them had to be taken offline for a period of time. The #allstate archive was taken offline during part of January and February, and thus resulted in the five-week period during which no tweets were captured.

What was interesting as well is that there were 40,258 unique users that generated the 68,370 tweets. This equates to an average of only 1.69 tweets per user. The Allstate corporate Twitter identity generated the largest number of tweets at 1,169, which only equated to 1.6% of the total tweets. Over 33,000 of the users made only one comment over the study time period, so the overall conversation was not dominated by just a few users. In fact, the top 100 users only accounted for 13.4% of the tweets, and the top 1,000 users only accounted for 29.2% of the total number of tweets. Even when looking at the traditional 80/20 rule (80% of the involvement comes from 20% of the participants), this particular data falls significantly short of this criteria. The top 20% of the users only account for 52.9% of the tweets.

This underscores one of the realities that underlie social media. The content is really driven by the community rather than specific users. Certainly, there are users that are more active than others, and this activity can be a source of interest for further investigation by a company. But overwhelmingly what social media brings is a sense of the feeling of the community. And if the same sentiment is being expressed by multiple individual users, then it may be something an insurer wants to pay attention to.

In addition to looking at the trends in tweets by month, we also have summarized the tweets by hour of the day to understand the most active times of day for the use of Twitter related to this archive. Figure 2 shows the summaries by hour in Eastern Time.

Figure 2: Number of Tweets by Hour of the Day



As can be seen from the chart above, the most active times for the use of Twitter for this archive were between 11:00 am and midnight Eastern Time. This time period will represent the most active time periods across most of the time zones across the United States. Therefore, this distribution of time periods appears to be reasonable. There is still activity outside of this time period, but the tweet activity is between one-third and two-thirds lower during the late night and early morning hours. This type of information might be helpful to those that are responsible for monitoring, analyzing, and responding to Twitter activity, especially if trends in the type of activity by time of day can be determined.

There are other types of general descriptive statistics that can be calculated based on the data. The goal in generally describing the data is to determine whether the data appears to be reasonable, to determine the applicability of the data for the intended purpose, and to identify any potential gaps or concerns with the data.

4. DATA PROCESSING

The analysis of social media data is heavily dependent on the ability to analyze text data. However, there are some unique considerations in the analysis of social media data that make it different than a normal text mining analysis. One major consideration is that social media data tends to be informal, so issues with misspellings and abbreviations will be a larger challenge. In addition, in the case of Twitter, there are certain symbols that actually do have a meaning and therefore extra care needs to be taken in cleansing the text.

Much of the work required to analyze social media data will be spent obtaining and preparing the data for analysis. This is not a trivial exercise, and the proper approach for a company will depend on the ultimate purpose of the analysis. While the purpose of this paper is not to provide a complete discussion on obtaining social media data, we have listed a few approaches here that the reader can pursue further. As described earlier, the source of the data for this analysis is twapperkeeper.com, which is a service that captures and archives Twitter posts. There are third party applications which can capture social media data from websites, and these applications appear to be both web-based services as well as stand-alone programs. Developers can also create computer programs which monitor and capture information from their social media sites. Programs also exist which scrape information from screens, and this can be applied to monitoring and collecting social media data from websites. Ultimately, companies will need to work with their information technology departments to determine the best approach for collecting and storing social media data.

The first step in analyzing the text data is to remove all the punctuation and symbols. This information generally does not add to the understanding of the text and will make it more difficult to decipher the words that are part of the tweet. This information includes single and double quotation marks, parentheses, punctuation marks, and stray symbols (dollar signs, stars, etc.). In the initial data cleaning, the signs that actually have meaning for Twitter (@, #) were retained.

Once the tweet has been cleansed of punctuation marks and symbols, then the tweet can be parsed into words. This parsing occurs by identifying spaces and using these spaces as the indication of one word ending and another word beginning. The number of potential words will depend on the source of the data. In the case of Twitter, posts are limited to 140 characters, so identifying up to 35 words for this analysis was sufficient. There are other sources of social media such as Facebook where one will need many more words than this to capture all the content.

At this point, the data is structured in a manner shown below:

Table 1: List of Words in Rows

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

Next, we want to determine the frequency of words present in these tweets. To do this, we stack all of the word columns into one column, and then summarize the word frequencies based on this combined column. This will change the data structure from tweets in rows to tweets in columns with one word per row. In stacking these words, care must be taken to maintain the Tweet ID and the word order, since this will be important later in the analysis. The structure of the data once this transformation is made is shown below:

Table 2: List of Words in Columns

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

This data can then be summarized by word to determine the frequency of words in the tweets. It is here that one will find many different types of words that may not be beneficial to the analysis, words such as "a, an, the, in, I, on, and of." Therefore, these words should be removed at this point so they do not unnecessarily slow down the analysis. The top 10 words in this analysis are shown below.

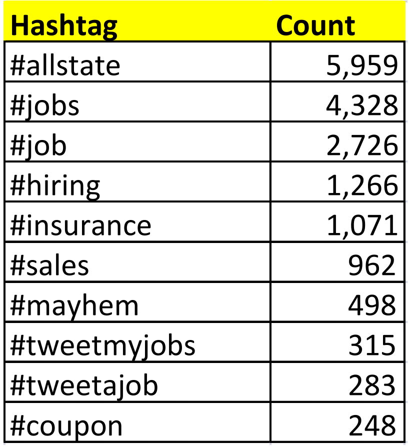
Table 3: Top 10 Keywords

Word	Word count	Pct of Words
allstate	70815	7.0%
insurance	16868	1.7%
Rt	9292	0.9%
jobs	6093	0.6%
commercials	5502	0.5%
arena	5327	0.5%
good	5132	0.5%
mayhem	5113	0.5%
job	4548	0.5%
like	4007	0.4%

The most prevalent keywords are not surprising. They are related to several different areas that are important to insurance companies. Obviously, the term "insurance" would be expected. "Jobs" appear to be an important topic in insurance company tweets, showing up in total over 10,000 times in the dataset. In addition, there are two words in the top ten related to Allstate advertising, including "commercials" and "Mayhem," which is one of the current advertising campaigns being run by Allstate. Lastly, there are two words related to insurance company slogans present in the top 10, "good" and "like."

In addition to the analysis of the frequency of words present in the tweets, because tweets use hashtags to identify keywords, an analysis of the keywords identified can also be undertaken. Based on the #allstate archive, the top 10 hashtags are:

Table 4: Hashtag Frequency



Hashtag	Count
#allstate	5,959
#jobs	4,328
#job	2,726
#hiring	1,266
#insurance	1,071
#sales	962
#mayhem	498
#tweetmyjobs	315
#tweetajob	283
#coupon	248

There are similar words present in the hashtags as there are in the keyword analysis. Five of the categories are related to employment, including either "jobs" or "hiring." Mayhem is also mentioned here as well. The difference in this list that stands out is the hashtag "coupon." After further investigation of these tweets, they appeared to be coupons offered by either Allstate as they participated in home shows or agencies offering coupons for local merchants.

A simple application of understanding the keywords that are present in tweets would be to set up rules that are triggered by certain keywords. For example, one the keywords present in the data is claims. This word could be a trigger for identifying insureds with questions or concerns about filing claims or the claims process. Another keyword is quote. This could be a trigger to identify potential customers who are looking for information on insurance prices.

There are two issues with the analysis of text data that can be corrected at this point. The first issue is misspelling, and the second issue is different cases or variations of the same word. Again, because of the informal nature of social media, misspellings are common. Also, as can be seen in the top ten words, there can be different tenses or cases of the same word, as with "job" and "jobs." In each of these cases, the desire in the analysis is to reflect the intent of the user. In order to do this, we want to either correct the spelling or make the different tenses consistent. Both of these issues can be addressed using the same approach.

Two techniques were used in this analysis to identify spelling and tense or case differences. One approach is the comparison of two strings by computing the Levenshtein edit distance (LED). The LED is defined as the number of insertions, deletions, or replacements of single characters that are required to convert one string to another. For example, the LED between "job" and "jobs" is 1,

since the strings would be equal by either adding an "s" to the first word or deleting the "s" from the second. Another approach to comparing strings is to calculate the generalized edit distance (GED). The GED between two strings (string1 and string2) is calculated as the minimum cost sequence of operations for constructing string2 from string1. In this calculation, different operations have different penalties associated with them. For example, inserting or deleting a character to create string1 incurs a penalty of 100, but the difference of a blank between the strings only incurs a penalty of 10 points. In the case of the example above, the generalized edit distance between "job" and "jobs" is 100 points, since inserting or deleting an "s" incurs a penalty of 100 points. [Note: The author used SAS™ for this analysis, and the specific SAS functions used were COMPLEV and COMPGED. A description of the COMPGED word operations and the points associated with each can be found at the following URL:
http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a002206133.htm.]

In the case of both distance calculations, the smaller the distance between two words, the more similar the two words are. The keywords that are identified can be compared to the remainder of the words to determine if there are misspellings, or if there are different tenses or cases of words that are present. Based on an investigation of the distances, a cutoff point can be selected to investigate further whether words can be considered as the same. Once these distances have been calculated, then the list of words can be edited to correct the misspellings and to make word variations consistent.

The last step in the data preparation is to add to Table 1 a set of indicators based on the identified keywords. Based on the keywords identified, an indicator can be added to the table that indicates the presence of a word in a particular tweet. For this analysis, there were 116 keywords identified, and so 116 indicators were added to the dataset that are either 0 or 1 depending on whether the word was present. The structure of the final table with tweets by row is shown below.

Table 5: Example of Structure of Final Table

Tweet ID	User	Tweet	allstate	insurance	commercial	company	…
19	@mosley	Allstate insurance company	1	1	0	1	…

5. TWEET ANALYSIS METHODOLOGY AND RESULTS

In the data exploration phase, keywords are identified in the data and adjustments are made to the data to prepare it for analysis. The purpose of the analysis of social media is to identify patterns and trends that are present in the information which may be of further use to the insurance company. To achieve this goal, we need to identify patterns and combinations of words that will indicate themes and ideas. One step in doing this is a simple correlation analysis, which will identify correlations between pairs of words. There are also two additional types of analyses that will be performed on the data. The first will be a clustering analysis which will group tweets based on their similarities or dissimilarities. The second will be an Association Analysis which analyzes the occurrence of specific words together.

5.1 Correlation Analysis

The correlation statistic used was a Cramer's V statistic for pairs of keyword indicators. Cramer's V indicates the level of association between two nominal variables. To calculate the Cramer's V statistic, assume a 2 x 2 matrix indicating the frequency of the combination of two words.

Table 6: Frequency of Word Combinations

Word 1 / Word 2	**0**	**1**
0	n_{00}	n_{01}
1	n_{10}	n_{11}

The notation n_{ij} represents the frequency of the combination of words in the dataset. For example, n_{00} counts the number of tweets where neither Word 1 or Word 2 were present. $n_{.j}$ is the total frequency for column *j*, while $n_{i.}$ is the frequency for row *i*. Given this two-by-two table structure, the formula for Cramer's V can be simplified to:

$$\text{Cramer's V} = \frac{n_{00}\, n_{11} - n_{01}\, n_{10}}{\sqrt{n_{0.}\, n_{1.}\, n_{.0}\, n_{.1}}} \qquad (1)$$

The result is a number between -1 and 1. A value of -1 indicates a perfect negative correlation, a value of 1 indicates a perfectly positive correlation, and 0 indicates no correlation. The top 20 combinations of words with the largest Cramer's V statistics are shown below.

Table 7: Top 20 Cramer's V Statistics

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	0.663
5	jobs	gravy	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

There are several categories of correlation that are apparent in this list. There are several correlations which are related to competitors, including State Farm, Progressive, and GEICO. There are also several pairs of words related to employment, including agency purchase, and jobs. There are also pairs of words related to characteristics of the company, which include "good hands," "personal financial," and "insurance company." Another category includes entertainment and other related items, such as "YouTube Jonas," "YouTube video," and "love basketball."

The YouTube and Jonas categories related to a public service campaign that was sponsored by Allstate in which the Jonas brothers participated. "Love and basketball" refers to a movie that Dennis Haysbert had a part in. Haysbert is now a spokesperson for Allstate, which is the connection to this dataset.

While the correlation is a simple approach that will begin to uncover combinations of words, it does not give a complete picture of the words and concepts that may be present in the tweet. The most obvious limitation is that only pairs of words are compared in this example. Since phrases can be up to 35 words long, understanding only the relationship between pairs of words will miss concepts that are present. In addition, it is difficult to understand how many records are impacted by certain pairs of correlated words because multiple combinations could be present in the same tweet. This problem is addressed by techniques that determine the presence of combinations of words and phrases, which are discussed in the remainder of this section.

5.2 Clustering

The next approach that can be applied to social media analysis is a cluster analysis. The clustering procedure is based on calculating distances between observations and is used to segment databases. Clusters are developed such that observations that are in the same cluster tend to be similar, and objects in different clusters tend to be dissimilar. The clusters developed in this paper use the Ward's Minimum-Variance method. Using this method, the distance between two clusters is calculated as the ANOVA sum of squares between the two clusters summed up over all the variables. The goal of each step of the process is to minimize the within cluster sum of squares.

This method was applied to the 116 keyword indicators which were identified in the data exploration phase. The results of applying this method to the #allstate archive resulted in 47 clusters, or 47 groups of observations that were combined based on their similarities. For each cluster, there are several ways the output can be viewed to attempt to understand what the results show. For each cluster, the percentage of tweets that contain each word is calculated, and from these percentages it can be determined which concepts are predominate within a particular cluster. In addition, the percentage of tweets in a cluster that contain a particular word can be compared to the overall percentage of tweets that contain that word. This will help analysts to see where words are showing up most, even if a keyword does not show up in a large percentage of tweets overall. The percentage of times a word shows up in a particular cluster can also be ranked across all clusters, which will also help the analyst see quickly where words are showing up most frequently.

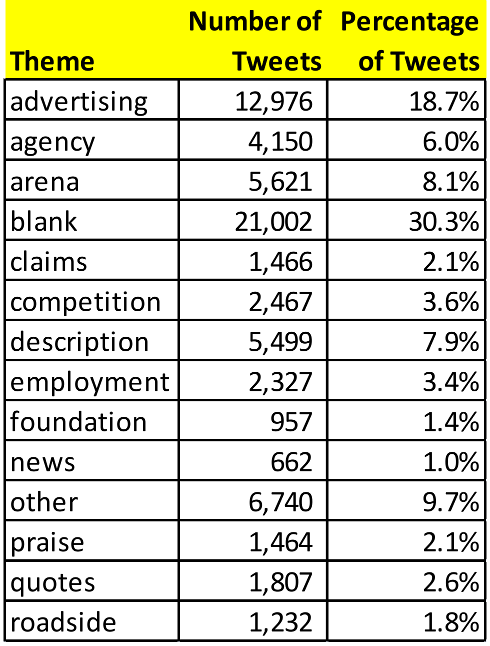
To determine how important a word is in a cluster, we calculate a ratio called cluster lift.

$$\text{Cluster Lift (word)} = \frac{\text{Percentage of tweets in a cluster that include word}}{\text{Percentage of all tweets that include word}} \qquad (2)$$

This calculation provides an indication of how much more likely a word is to be in a particular cluster than it is present in the dataset overall. The cluster lifts are shown in Exhibit 1. This is shown for a selection of the keywords. We have highlighted cells with a cluster lift greater than 4.0 to show the keywords that are prevalent in each cluster (4.0 was chosen for demonstrative purposes – the proper level of significance for a particular analysis should be determined by the reader). For example, in cluster 18, the word "mayhem" has a cluster lift greater than 15, and the word "guy" has a cluster lift of 5. Therefore, it can be reasonably concluded that tweets in this cluster have something to do with Allstate's commercial personality. In cluster 12, the term "good" has a cluster lift of 12.8, and "hands" has a cluster lift of 19.5. This cluster thus has tweets related to the slogan associated with Allstate.

Exhibit 2 shows the keywords for each cluster that are greater than the 4.0 threshold. Based on the keywords present, the clusters can be grouped into general themes. In reviewing the keywords from the cluster analysis, the tweets were grouped into the themes shown in the table below.

Table 8: Key Themes



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%

There are several key themes which are present in the analysis. Ignoring the blank and other categories for a moment (we will come back to them), the largest percentage of tweets are related to advertising. More specifically, most of these tweets were discussing the Allstate commercials. While there were some reactions on both sides, the majority of the tweets related to the commercials were positive, with words being used like "funny," "like," and "love." The next largest category was a category associated with the Allstate arena, referencing upcoming or recent concerts and sporting events there. There were also a significant number of tweets related to employment with Allstate, many describing opportunities as an agent, but some referring to employment opportunities with the company. Many of these showed up in the categories labeled "agency," "employment," and "description." More detailed analyses should be performed on each of the significant themes to determine the types of things, both positive and negative, that are being said about the particular area.

Once these general themes or concepts are extracted from the results, the right areas within the company can be brought in to discuss what to do with the results. For example, given the significant number of tweets regarding the advertising, the marketing department may be able to take feedback from the comments posted by users to improve or build upon an advertising campaign. There was also a theme related to claims. Claim department executives would most likely be interested in hearing this unsolicited feedback that is being provided not to the company, but to a community of millions of users. This information could be used by the claims department in a number of ways, including improving the claims process and addressing complaints about claims handling.

As mentioned, the largest two categories are the missing and the other categories. The cluster with no keywords occurred because there were no keywords in the tweets that met the threshold that was set, which was four times greater than average. For the clusters in the other categories, there were a collection of words that related to a number of different things, but no general theme was identified. This will generally be the case, especially when dealing with social media data. There will be a collection of posts or data elements that a text mining process may have some difficulty classifying. Relating back to the Pear Analytics study, these could be tweets that fall into the pointless babble, self-promotion, or spam categories. There are a couple of ways that these issues can be addressed. The first is to try and do a more detailed analysis on these specific clusters, analyzing the presence of keywords or running a separate cluster analysis on this subset. Another approach is to use a rank of cluster lifts to find the most prevalent words in the cluster, whether or not they meet the standard of four times the average.

When producing this ranking, which is shown in Exhibit 3, it can be seen that several different concepts are present in cluster 21, which was the cluster that had no keywords that met the threshold. The five words that had the largest cluster lifts are, in order: Northbrook, personal, rep, claim, and April. As can be seen, these tweets represent a series of different ideas. Northbrook is the location of the Allstate corporate headquarters, and there also appears to be some references to claim representatives as well. A combination of investigating some of the tweets as well as applying some of these word and concept analyses to the tweets specifically within that cluster will help uncover more detail in the cluster and potentially extract more information from those tweets.

It is also here that we can begin to see the reason for the increase in tweets in March, 2011. The increase essentially came from several categories of tweets. There was a significant increase during the time period of tweets related to employment and agency opportunities with Allstate, and there was also an increase in the number of tweets related to current and future events at the Allstate arena. There also appeared to be an increase in the number of tweets soliciting customers to receive quotes for insurance during this period.

5.3 Association Analysis

Cluster analyses are helpful in that they create groups of data points that have a relationship with each other, and these relationships can be examined in more detail to discover underlying concepts in the groups. The cluster analysis does have some disadvantages, though. The first is that each data point can belong to one and only one group. For examples like this one where Twitter messages are by definition short, this will not be a huge disadvantage because most tweets will only be focused on one concept. However, for social media outlets without such limitations, like Facebook or a blog, you may have a case where multiple concepts and themes will be present in each data point. In this situation, it will be difficult to assign data points to one and only one group. To address this, an analysis that highlights combinations of words regardless of where they occur can also help an analyst understand the key and important concepts in a body of text data.

One of the techniques that can be applied to highlight combinations of words is association analysis. Association analysis has its background in market basket analysis. It is used in retail environments, such as grocery stores or pharmacies, to identify items that tend to be purchased together. This allows stores to optimize the layout of their store and potentially increase sales by cross-selling or up-selling customers. This analysis determines the likelihood of a combination of items occurring together as well as a confidence around the projection. Ultimately, the association analysis produces a set of if-then rules (if item A is present in a transaction, then item B will be present as well), and the lift associated with the rule.

There are several calculations that are made as part of an association analysis to determine the strength of relationships. The support is a measure of how often items occur together.

$$\text{Support} = \frac{\text{Transactions that contain items A \& B}}{\text{All transactions}} \qquad (3)$$

Confidence measures the strength of the association by measuring how often item B is present when item A is present.

$$\text{Confidence} = \frac{\text{Transactions that contain items A \& B}}{\text{Transactions that contain item A}} \qquad (4)$$

The Expected Confidence is the proportion of items that satisfy the right side of the if-then association rule. This provides the expected presence of Item B if there was no relationship between Items A and B.

$$\text{Expected Confidence} = \frac{\text{Transactions that contain item B}}{\text{All transactions}} \qquad (5)$$

The lift can then be calculated as the ratio of the confidence (4) to the expected confidence (5). The higher the lift, the more the presence of Item B is influenced by the presence of Item A.

Applied to text mining, transactions would be the text field, and the items would be the words themselves. More specifically, in this analysis, a transaction would be a tweet, and the items would be the words in the tweet.

There can be literally thousands of rules generated from an association analysis, and for this example the author calculated 2,000 rules. The lift of the association rules ranged from a high of 140.21 to a low of 11.6. This means that we could calculate more than 2,000 rules and still generate rules that have a significant lift. Exhibit 4 shows an extract of the rules that were generated as a result of this analysis. As indicated above, many of the rules were associated with jobs and employment, so these were excluded those from this extract. The extract shows the statistics described above and a description of the rules. For example, notice Rule Index 281. The left side of this rule is TV, and the right hand of the rule is “Mayhem & ad.” The expected confidence is .9%, meaning that Mayhem & ad only appear in only .9% of the tweets. However, when the word TV is present, Mayhem and ad occur in 50.9% of the tweets, as shown in the confidence percentage column. The lift for this combination is over 55. As a result of reviewing these rules, subjects of tweets can be identified. Tweets that satisfy these rules can be identified and the sentiments that are being expressed related to these topics can be explored further.

The results of an association analysis can also be viewed using a link graph. A link graph displays the rules by using nodes and links. The size of the node varies based on the number of transaction counts (in this case, tweets) that the rule relates to, and the color and thickness of the line varies based on the strength of the relationship. Exhibit 5 shows an example of the entire link graph based on the analysis of the tweets. As you can see from the graph, there is a large mass of nodes and links, and then a number of nodes and links that surround the big mass. The big mass of links in the middle relates to a series of job, employment, agency, and financial opportunities with Allstate or its agencies. There was a significant amount of activity during this time period related to these opportunities, and tweets with the same or similar content were sent many times over the time period. Obviously, more detailed analysis would need to be done on this group to determine the value of the information here.

Exhibit 6 shows the link graph zoomed in on the top left. This zoom represents a collection of words related to "good," which is the most prevalent word in this set of rules. Looking at this group of tweets, there are actually a couple different themes being represented here. The first is the Allstate slogan, "You're in Good Hands." The other concept is the State Farm slogan, "Like a Good Neighbor, State Farm is There." So even though these tweets are related to the hashtag #allstate, there are a number of references to their competitor's slogan in these tweets. Many of these tweets were related to tweets of customers expressing their opinions regarding the different companies, and some were even related to jokes that were circulating around Twitter regarding the two insurance companies.

Exhibit 7 shows a zoom on the right side of the link graph. Here, there are several association rules created related to competitors (State Farm, GEICO, Progressive), and several related to the TV ads (Mayhem, TV, ad). In addition, there were a couple of other rules that highlight specific areas of interest related to Allstate. One was related to the Allstate "X The TXT' campaign, which involved the Jonas Brothers. This was a campaign to highlight the dangers of texting while driving. Many tweets were sent that expressed a positive reaction to this campaign. There was also a rule related to "$1 billion." This was a reference to the purchase price that Allstate paid to acquire Esurance. This news was tweeted about pretty heavily for a short time period when the acquisition was made public.

Again, the purpose of the association analysis is to highlight words and concepts that are coming through in the data. Once these areas have been highlighted, the company can decide how this information should be disseminated and applied. For example, the mixing of slogans might highlight an opportunity to more clearly define the brand, while the interest in jobs and employment with Allstate might help develop a targeted recruiting campaign.

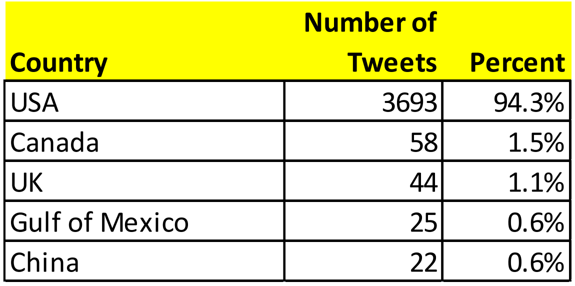
Although not covered in this paper, as an extension to this analysis, sequencing can also be done. A sequence analysis looks not only at the association of words within phrases, but also analyzes the order of the words within the phrases. This analysis then provides the likelihood of particular word orders and phrases, which can potentially give further insight into the content, especially when re-ordering words may change the meaning of a phrase. The investigation of sequence analysis is left to the reader.

5.4 Geographical Information

The Twitter location information that is available from mobile Twitter applications can save where the user was when the tweet was sent. The user generally has the option to turn location services on or off. In this analysis, 3,918 tweets have location information saved with them, which is 5.7% of the tweets. Associating the tweet with the location can have obvious applications for an insurance company. There could be claim implications, especially if a customer is somehow using social media right after a claim happens. Also, if there are multiple users within a certain geographical area that are tweeting about the same issue (premium quotes, for example), this may identify a concern or an opportunity for the company.

Exhibit 8 shows a map of the locations of the users for the tweets where the location was populated. The top five countries are shown in Table 9. As to be expected, most of the tweets originated from the United States and Canada. There were also a small percentage of tweets that were sent from the United Kingdom and China, and even some from the Gulf of Mexico.

Table 9: Top 5 Countries



	Number of	
Country	**Tweets**	**Percent**
USA	3693	94.3%
Canada	58	1.5%
UK	44	1.1%
Gulf of Mexico	25	0.6%
China	22	0.6%

We also focused our attention on specific states that were generating lots of tweets. There were a large percentage of tweets in several states including Virginia, Washington, and Illinois that were related to financial representative and agency opportunities with Allstate. Many of the tweets from Illinois were users talking about and checking in from events that were going on at the Allstate Arena. There were also a few tweets related to roadside assistance help in several states. A large number of the location tweets were related to Foursquare check-ins, which allows users to check in at particular locations, and these check-ins can then be passed through Twitter.

6. CHALLENGES WITH SOCIAL MEDIA ANALYSIS

While the value that can be gained from analyzing social media data is great, there are challenges associated with social media analyses which will require further exploration. One of the first challenges will actually be accessing and collecting this information. As we discussed earlier, there are applications available which will allow companies to begin collecting and analyzing social media data, and companies may also have the ability to build internal programs that do this. The key is to make sure that the data is being collected in a consistent and complete matter, and that it is easily accessible for analysis.

There are also challenges in analyzing social media as it relates to analyzing text data. One of these challenges relates to the context. There are many times when a Facebook post or a tweet is simply a response to another post or tweet. Depending on how the user is responding or what a company may be tracking, the information that is being responded to may or may not be available. To the extent that it is not available, this creates a challenge for the analysis to understand exactly what this information means. If it is available, the challenge becomes connecting the right set of social media data together to be able to understand the broader context of a conversation. This is not a trivial exercise, and there is still emerging work being done to improve the analytics in this respect.

In this analysis, we are assuming that each tweet has equal weight. However, there are reasons that one might want to weight tweets differently. One reason is because users have different numbers of followers, and a tweet from a user with 1,000 followers is likely to be seen by more users that a tweet from a user with 100 followers. So a tweet could be given a higher weight given the influence of the person sending the tweet. Another reason for giving a tweet a higher weight may be the fact that it is a retweet. If it is a comment that other users are agreeing with, it can spread faster and influence more users. So the number of followers and number of retweets should be considered in the weighting of the tweets.

The focus of this paper has been about determining the subject matter associated with a tweet. Another challenge in terms of the analysis of social media data is understanding customer sentiment. Words in a tweet are simply that, words. They do not carry the normal emotion that is present in a face-to-face conversation, in which case the listener could detect happiness, sadness, sarcasm, etc. There are things that users attempt to do to try and convey different emotions (smiley face, sad face, "lol," TYPING IN ALL CAPS, etc.), but even when a person reads an electronic communication they might get the sentiment wrong. Depending on the forum being analyzed, there are a number of understood rules that communicate different things, such as changing the font color to indicate sarcasm. While this analysis focused on the words being used, to understand sentiment would require a more thorough investigation into the ways that users communicate sentiment, and then attempting to capture those sentiments within the data in a structured way. This is another challenging area where there is still work to be done.

Another issue inherent in social media data arises from the fact that social media data is unfiltered. There are no system edits that ensure the social media data that was captured is accurate, and this may result in false information and statements that are driven by pure emotion rather than fact. This will make the process of sifting fact from fiction a delicate one. However, companies can guard against this by not simply accepting all the things that come out of a social media analysis, but filtering the results through their overall understanding of the business. Also, generally for a trend to show up significantly in a social media analysis, there has to be more than a few users making statements, but multiple users expressing similar sentiments that cause it to rise to the top. Ultimately, it will be very important for the analyst to consider the source of the data in the interpretation of the results.

Typical actuarial predictive modeling analyses are based on historical data that is usually at least three months old and could be at least a year old. Also, typical actuarial predictive analyses are repeated relatively infrequently. Generally, it is at least a year until an analysis is repeated, and could be even longer than that. In order to be able to apply the results of a social media analysis in a timely manner, the analysis will need to occur frequently, and as close to real-time as possible. Trending topics can literally begin in an instant and can become widespread very fast, and if the analysis occurs too long after the topic is trending, it may be too late for the company to do anything useful about it. Therefore, the analysis will need to be automated such that it can be updated quickly and the results reviewed in a timely manner.

Lastly, the world of social media is not limited to those that use the English language. This will especially be true for companies that have an international presence, and even true for companies that function solely in the United States. The concepts discussed in this paper apply generally to other languages. However, if a company has social media data which includes information in multiple languages, the differences in the languages will necessitate at least a separate initial analysis. This analysis could then be combined later if it is possible to translate the words and sentiments into one language.

7. APPLICATIONS OF SOCIAL MEDIA ANALYTICS

As can be seen in the discussion above, there are a number of things that come out of the analysis of social media. A number of different techniques can be applied to understand the words and combination of words present in each tweet, and how active and popular these particular topics and conversations are. Ultimately, the result of this data-mining exercise will be an identification of trends in the social media conversation, and as a result of the identification of these trends, there are a number of ways that an insurance company can apply this information.

One area of application is in customer service. Some of the tweets in the analysis were related to claims, and depending on the content of the analysis, this information could be used to address potential concerns or questions regarding the claim process. There could also be questions raised related to policy information or provisions in a social media forum, and identifying and proactively addressing these issues gives the insurer an opportunity to provide superior customer service to their policyholders.

Another application of social media analytics would be a better understanding of customer sentiment about the company. One example of this is the customer reaction to company advertising campaigns. As customers provide feedback on advertising and commercials, for example, companies can use this raw, unsolicited feedback to make their marketing programs more effective. Also, if customers are particularly happy or unhappy with a company about a particular issue, it provides the company with an opportunity to attempt to proactively address this issue before it takes on a life of its own.

Social media analytics would also allow the insurance company to gather competitive intelligence from several different perspectives. In this analysis, there were some users that were very vocal about their preferences of one insurance company over another. While an entire strategy cannot be based on the feelings of a few current or potential customers, if customer sentiment begins to build for or against the company or a competitor, this can be understood and the company can react to it. Also, in this analysis, there were obviously some users who may have been employees or agents of other companies that were soliciting customers. These trends can also be identified and monitored, and could potentially provide insight into competitive issues.

From a broader perspective, the use of social media can be used to identify broader trends in the market that the company may be able to take advantage of. One example of this might be an influx of social media data that suggests more people are looking for quotes or shopping for insurance in a particular area, or identifying concerns with finding affordable insurance in a particular area. Again, this could be brought forward to the right area within a company and proper steps taken to respond to these market trends.

8. CONCLUSIONS

As can be seen all around us, the use of social media has grown significantly, and has transformed the way that people interact. This growth in social media has led to an increase in the amount of information that is being generated, and this information provides insight to companies, including insurers, related to their business. Analytics can be applied to social media to identify key words and phrases that are being expressed, and these findings can be used by insurers to assist in managing their business, and to interact more effectively with current and potential customers.

Acknowledgment

The author acknowledges Gary Wang and Nick Kucera for their review and for comments that improved the paper.

9. REFERENCES

[1.] Banking.com Staff, "Social Media Statistics: By-the-Numbers, January, 2011," Banking.com. 24 Jan. 2011, accessed 22 Aug. 2011, http://www.banking2020.com/2011/01/24/social-media-statistics-by-the-numbers-january-2011-part-ii/>.

[2.] Brown, Danny, "52 Cool Facts About Social Media," Danny Brown. 3 Jul. 2010, accessed 22 Aug. 2011, < http://dannybrown.me/2010/07/03/cool-facts-about-social-media/>.

[3.] Browne, Sean, "Statistics: Social Networks will Receive 11% of Online Ad Spending in 2011," *Wealthvest Marketing*, 20 Jan. 2011, accessed 22 Aug. 2011, < http://www.wealthvest.com/blog/2011/01/20/statistics-social-networks-will-recieve-11-of-online-ad-spending-in-2011/>.

[4.] Customer Respect Group, "Social Eyes: The Insurers' View of Social Media," Volume 1, Number 5. July, 2011.

[5.] Francis, Louise A., "Taming Text: An Introduction to Text Mining," Casualty Actuarial Society *Forum,* Winter 2006, pp. 51–88, http://www.casact.org/pubs/forum/06wforum/06w55.pdf.

[6.] Kelly, Ryan, "Twitter Study Reveals Interesting Results About Usage—40% is 'Pointless Babble,'" Pear Analytics blog post, http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results-40-percent-pointless-babble/.

[7.] "Twitter Study–August 2009," August 12, 2009, San Antonio, TX: Pear Analytics, http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

Biography of the Author

Roosevelt C. Mosley, Jr. is a principal with Pinnacle Actuarial Resources, Inc. Roosevelt has 18 years of experience in the property and casualty actuarial field, including over a decade of experience in the application of advanced analytic techniques to insurance companies. Roosevelt is a Fellow of the Casualty Actuarial Society and a Member of the American Academy of Actuaries. Roosevelt's experience in the area of insurance analytics includes rating, underwriting, claims, and marketing.

Tweet Analysis
Cluster Results

Exhibit 1

Cluster	insurance	rt	commercial	arena	good	mayhem	job	like	commercial	hands	car	company	auto	financial	guy	lol	agent
1	0.519	7.879	1.778	0.000	0.741	2.333	0.000	17.635	0.054	0.722	0.627	0.000	0.000	0.000	2.805	2.426	0.216
2	0.300	0.082	0.000	0.000	0.000	0.000	6.631	0.010	0.000	0.043	0.097	0.022	0.036	0.113	0.038	0.014	24.915
3	1.471	0.572	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.142	0.259	5.575	20.240	0.000	0.000	0.000
4	0.002	7.879	0.598	0.000	0.000	0.399	0.325	0.000	0.000	0.252	0.420	0.118	0.123	0.453	0.780	0.000	0.237
5	0.169	1.697	2.118	0.018	0.047	0.149	0.055	1.287	3.412	0.175	0.289	0.000	0.046	0.000	2.702	25.118	0.105
6	0.000	0.000	0.025	0.000	0.000	0.000	0.315	0.000	0.000	0.194	1.097	0.121	0.308	0.000	0.456	0.000	0.050
7	0.313	0.000	1.335	0.063	0.047	1.990	0.081	17.635	0.000	0.641	0.407	0.083	0.015	0.000	2.384	0.343	0.138
8	0.000	0.000	0.000	0.000	0.000	0.000	0.055	0.000	0.000	0.054	0.122	0.056	0.000	0.000	0.222	0.000	0.000
9	5.215	0.408	0.000	0.000	0.222	0.000	0.641	0.057	0.000	0.000	0.714	13.089	22.112	0.000	0.000	0.000	2.951
10	5.196	0.484	0.014	0.000	0.036	0.000	1.555	0.065	0.007	0.014	0.000	19.571	0.008	5.789	0.017	0.019	3.778
11	4.715	0.183	0.037	0.000	0.120	0.045	0.000	0.103	0.055	0.000	2.054	0.118	13.177	0.000	0.000	0.000	0.147
12	0.000	1.447	0.145	0.000	12.888	0.212	0.079	3.092	0.420	19.540	0.221	0.020	0.022	0.007	0.506	2.517	0.144
13	0.036	0.791	0.177	3.909	0.190	0.268	0.411	0.366	0.264	0.000	0.229	0.000	0.459	0.071	0.318	0.087	0.175
14	3.824	0.274	0.174	0.000	0.234	0.000	0.162	0.132	0.208	0.013	22.076	0.524	1.056	0.042	0.345	0.274	0.691
15	3.016	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	9.967	0.000	10.148	0.000	0.000	0.000	0.000
16	0.040	7.757	0.396	0.000	13.737	5.165	0.307	0.137	0.222	0.151	0.171	0.000	0.000	0.000	7.208	1.266	0.098
17	0.000	0.027	0.000	10.828	0.114	0.000	0.027	0.280	0.000	0.000	0.000	0.000	0.000	0.000	0.095	0.087	0.000
18	0.000	0.510	0.000	0.000	0.073	15.496	0.024	0.052	4.322	0.012	0.354	0.012	0.000	0.037	5.055	0.969	0.075
19	0.263	1.080	0.215	0.000	0.231	1.639	0.095	17.635	19.066	0.141	1.061	0.000	0.000	0.000	3.587	3.079	0.000
20	0.000	6.399	0.000	0.000	13.737	0.000	0.000	17.466	0.000	0.000	17.084	0.000	0.000	0.000	0.073	0.000	0.000
21	0.000	0.000	0.000	0.000	0.000	0.000	1.721	0.000	0.000	0.001	0.000	0.226	0.000	1.874	0.610	0.000	0.000
22	0.248	2.412	0.434	0.000	12.896	0.000	0.000	12.476	0.000	1.994	8.410	0.000	0.451	0.000	1.406	1.025	0.172
23	2.929	0.614	0.818	0.000	0.190	0.350	0.000	0.532	0.479	0.049	1.997	0.279	3.611	0.000	0.548	0.978	0.414
24	0.356	1.074	0.290	0.000	13.113	0.352	0.000	15.831	0.000	3.553	3.261	0.460	0.000	0.000	0.261	2.854	0.576
25	0.379	0.849	0.110	0.013	0.046	0.000	0.133	0.000	19.066	0.056	0.338	0.000	0.032	0.000	2.517	0.000	0.036
26	0.000	0.641	0.020	0.000	0.000	0.000	0.190	0.028	0.000	0.031	0.880	0.613	5.607	0.033	0.000	0.120	0.081
27	0.000	7.879	0.000	13.135	0.066	0.000	0.038	0.034	0.000	0.000	0.000	0.000	0.000	0.000	0.044	0.192	0.000
28	5.215	0.494	0.231	0.013	0.000	0.000	1.967	0.021	0.009	0.047	0.000	0.000	0.000	0.129	0.415	0.218	1.575
29	0.046	0.669	12.772	0.000	0.048	6.392	0.162	0.490	0.106	0.023	0.181	0.000	0.006	0.000	3.791	1.176	0.015
30	4.087	0.234	0.438	0.000	0.027	0.010	0.497	0.081	0.579	0.064	0.015	0.093	22.112	0.041	0.045	0.133	0.167
31	5.215	0.124	0.000	0.000	0.432	0.000	0.000	0.000	0.037	0.000	17.696	0.159	7.211	0.243	0.000	0.000	0.000
32	0.402	2.072	2.735	0.000	0.346	0.000	0.062	0.750	4.143	0.462	0.139	0.000	0.000	0.000	7.267	5.854	0.000
33	0.029	0.077	0.053	0.073	0.019	6.543	0.000	0.122	0.662	0.000	0.123	0.056	0.031	0.029	0.096	0.105	0.106
34	4.874	0.110	0.060	0.000	0.128	0.000	0.000	0.330	0.000	0.000	20.219	4.819	12.193	0.000	0.000	0.117	0.000
35	0.395	1.441	1.631	0.000	0.057	0.664	0.065	1.641	0.660	0.241	0.164	0.000	0.055	0.000	0.378	3.331	0.000
36	5.215	1.828	0.247	0.000	13.623	0.086	0.876	1.413	0.105	14.628	1.220	0.670	2.993	0.000	0.444	1.110	0.350
37	0.401	1.556	0.164	0.000	0.070	0.119	0.203	0.136	0.244	0.100	0.679	0.000	0.340	0.000	0.235	0.580	0.065
38	0.002	0.000	0.000	13.135	0.082	0.005	0.028	0.267	0.000	0.007	0.062	0.007	0.000	0.000	0.065	0.327	0.000
39	5.215	1.602	5.247	0.000	0.114	15.303	0.000	1.829	2.136	0.081	3.664	0.671	1.009	0.000	4.954	0.834	0.210
40	5.170	0.310	0.056	0.000	0.060	0.068	0.000	0.154	0.000	0.000	0.000	0.177	0.193	0.270	0.000	0.000	0.221
41	0.000	0.000	0.411	0.338	13.737	0.538	0.816	0.976	0.294	0.000	0.369	0.078	0.114	0.027	0.561	0.970	0.326
42	0.208	1.050	12.308	0.000	0.119	3.780	0.069	0.580	0.132	0.000	0.420	0.000	0.000	0.000	2.423	3.303	0.000
43	0.780	0.447	0.000	0.000	0.071	0.000	0.000	0.000	0.000	0.000	0.114	0.104	1.824	2.872	0.000	0.000	0.000
44	1.134	0.690	2.199	0.000	0.166	0.234	0.120	0.586	0.230	0.000	1.067	0.061	0.534	0.000	1.110	0.379	0.230
45	3.187	0.297	0.000	0.026	0.709	0.000	3.853	0.000	0.000	0.000	0.044	4.935	0.044	0.901	0.000	0.050	0.854
46	0.069	3.204	0.106	0.000	0.023	0.077	0.459	0.000	0.000	0.032	0.328	0.033	0.037	0.171	0.228	0.042	0.335
47	0.000	0.000	0.028	0.029	0.000	0.000	0.000	0.039	0.000	0.043	0.146	0.000	1.264	0.045	0.050	0.166	0.445

Tweet Analysis
Cluster Keywords

Exhibit 2

Cluster	Number of Tweets	Keywords
1	352	rt like look please foundation
2	1,823	jobs job agent gravy hiring exclusive start states
3	468	auto financial home esurance quotes online news billion business answer
4	3,416	rt
5	1,453	lol black ***
6	503	going after
7	1,464	like look
8	725	tickets quote may rosemont people save year_w first
9	309	insurance company auto free geico more quotes claim progressive quote business
10	2,702	insurance company financial personal rep agency purchase
11	344	insurance auto free home rates best quote online life save safe drivers
12	2,994	good hands *** always roadside
13	289	time great show night year_w after concert last
14	1,466	car coverage accident
15	268	car auto free geico today rates esurance best great progressive coverage save look bad
16	258	rt good mayhem guy more progressive money team bad last roadside
17	1,446	arena chicago tickets show may rosemont tonight live glee concert
18	1,685	mayhem commercial guy hot hilarious
19	416	like commercial man lmao off night voice look
20	314	rt good like car state farm girl watch work neighbor statefarm
21	21,002	
22	147	good like car state farm girl watch work neighbor
23	796	state geico farm rates progressive life april bad
24	88	good like watch work neighbor bad statefarm
25	2,089	commercial funny lmao voice
26	627	auto home after safe win
27	1,045	rt arena chicago show off tonight live office glee concert first
28	4,150	insurance agency purchase
29	3,416	commercials mayhem love hilarious
30	1,516	insurance auto online life states motorcycle
31	509	insurance car auto state free geico today rates quotes check quote coverage online save drivers
32	635	commercial guy lol love man black *** voice basketball always
33	720	mayhem check youtube video jonas
34	214	insurance car company auto state geico farm today quotes coverage may life accident
35	1,214	man black voice
36	362	insurance good hands may life *** always
37	390	check help news driving team safe drivers foundation
38	2,841	arena tickets show may tonight live glee concert
39	241	insurance commercials mayhem guy
40	229	insurance free quotes coverage life year_w motorcycle
41	777	good team roadside
42	578	commercials funny lmao
43	194	sales esurance more online claims billion answer money
44	331	free geico esurance check progressive save statefarm
45	504	company sales hiring agency service bad
46	605	today foundation
47	455	free help service roadside

Tweet Analysis
Top 5 Keywords by Cluster

Exhibit 3

	Rank				
Cluster	**1**	**2**	**3**	**4**	**5**
1	foundation	please	like	look	rt
2	exclusive	agent	gravy	start	jobs
3	answer	esurance	billion	financial	home
4	rt	game	never	foundation	check
5	lol	***	black	commercial	still
6	after	going	game	tonight	still
7	like	look	better	please	really
8	people	quote	year_w	tickets	save
9	progressive	quotes	quote	auto	more
10	company	purchase	rep	personal	financial
11	quote	free	auto	life	best
12	hands	good	always	roadside	***
13	last	night	year_w	show	after
14	car	accident	coverage	drivers	insurance
15	rates	progressive	save	today	geico
16	money	progressive	roadside	good	last
17	rosemont	chicago	tickets	arena	glee
18	mayhem	guy	hot	commercial	hilarious
19	commercial	like	look	night	voice
20	neighbor	girl	watch	work	statefarm
21	northbrook	personal	rep	claim	april
22	neighbor	farm	state	work	good
23	farm	state	bad	april	geico
24	statefarm	neighbor	like	good	bad
25	commercial	funny	lmao	voice	hilarious
26	home	win	auto	after	safe
27	chicago	tonight	arena	office	concert
28	purchase	agency	insurance	motorcycle	home
29	commercials	hilarious	mayhem	love	guy
30	auto	states	online	motorcycle	life
31	quotes	rates	drivers	car	free
32	basketball	love	guy	black	***
33	video	youtube	jonas	check	mayhem
34	coverage	farm	geico	state	accident
35	man	voice	black	never	lmao
36	always	hands	good	***	may
37	driving	safe	foundation	team	help
38	glee	arena	concert	tickets	live
39	mayhem	commercials	insurance	guy	best
40	quotes	coverage	year_w	motorcycle	life
41	good	team	roadside	bad	look
42	funny	commercials	lmao	mayhem	lol
43	online	sales	esurance	money	answer
44	geico	progressive	statefarm	save	esurance
45	sales	hiring	bad	agency	company
46	today	foundation	check	great	rt
47	free	help	service	roadside	quote

Tweet Analysis
Association Rules

Exhibit 4

Expected				Transaction				Rule
Confidence(%)	**Confidence(%)**	**Support(%)**	**Lift**	**Count**	**Rule**	**Left Hand of Rule**	**Right Hand of Rule**	**Index**
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	78
0.754	81.043	0.682	107.516	466	neighbor ==> there & good	neighbor	there & good	110
0.771	79.174	0.618	102.644	422	neighbor & good ==> there & like	neighbor & good	there & like	134
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	281
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	88.285	0.618	53.855	422	there & neighbor ==> like & good	there & neighbor	like & good	296
1.041	52.679	0.777	50.620	531	ad ==> tv & mayhem	ad	tv & mayhem	311
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	331
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
2.355	89.597	0.618	38.045	422	neighbor & like & good ==> there	neighbor & like & good	there	590
1.527	57.837	0.853	37.886	583	ad ==> tv	ad	tv	595
2.355	88.889	0.632	37.744	432	neighbor & like ==> there	neighbor & like	there	602
2.355	87.430	0.682	37.125	466	neighbor & good ==> there	neighbor & good	there	648
0.826	30.318	0.809	36.727	553	state ==> farm & car	state	farm & car	665
2.670	98.050	0.809	36.727	553	farm & car ==> state	farm & car	state	666
2.122	77.138	2.059	36.346	1,407	state ==> farm	state	farm	673
2.670	97.034	2.059	36.346	1,407	farm ==> state	farm	state	674
2.670	96.434	0.950	36.121	649	insurance & farm ==> state	insurance & farm	state	702
2.355	83.130	0.700	35.299	478	neighbor ==> there	neighbor	there	738
2.128	58.010	0.700	27.258	478	progressive ==> geico	progressive	geico	955
0.973	23.359	0.886	23.999	605	love ==> basketball	love	basketball	1141
3.791	90.977	0.886	23.999	605	basketball ==> love	basketball	love	1142
0.733	14.359	0.697	19.582	476	financial ==> esurance & answer	financial	esurance & answer	1349
2.355	38.571	0.632	16.378	432	like & good ==> there	like & good	there	1493
1.639	26.849	0.632	16.378	432	there ==> like & good	there	like & good	1494
5.676	90.558	0.618	15.954	422	there & neighbor & good ==> like	there & neighbor & good	like	1550
0.682	10.882	0.618	15.954	422	like ==> there & neighbor & good	like	there & neighbor & good	1549
5.676	90.377	0.632	15.922	432	there & neighbor ==> like	there & neighbor	like	1553
0.700	11.140	0.632	15.922	432	like ==> there & neighbor	like	there & neighbor	1554
5.676	88.368	0.689	15.568	471	neighbor & good ==> like	neighbor & good	like	1573
5.676	84.522	0.711	14.891	486	neighbor ==> like	neighbor	like	1613
5.676	83.883	0.632	14.778	432	there & good ==> like	there & good	like	1629
0.754	11.140	0.632	14.778	432	like ==> there & good	like	there & good	1630
0.632	8.470	0.618	13.396	422	good ==> there & neighbor & like	good	there & neighbor & like	1725
0.700	9.354	0.682	13.369	466	good ==> there & neighbor	good	there & neighbor	1728
7.141	91.081	0.777	12.754	531	tv & ad ==> mayhem	tv & ad	mayhem	1766
0.853	10.883	0.777	12.754	531	mayhem ==> tv & ad	mayhem	tv & ad	1765
0.950	11.642	0.849	12.256	580	good ==> like & hands	good	like & hands	1886

Exhibit 5

Association Analysis
Link Graph

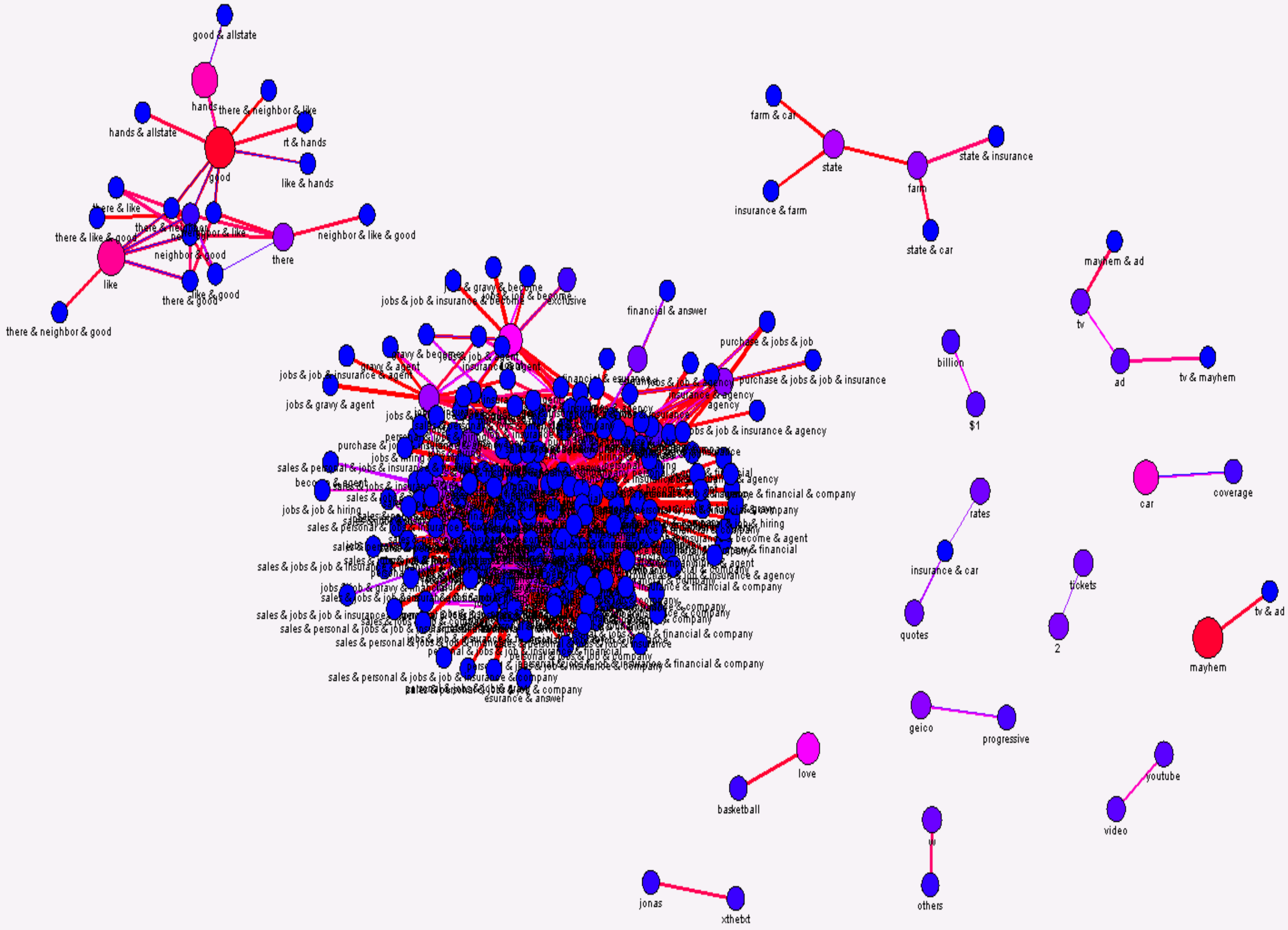


Exhibit 6

Association Analysis
Link Graph

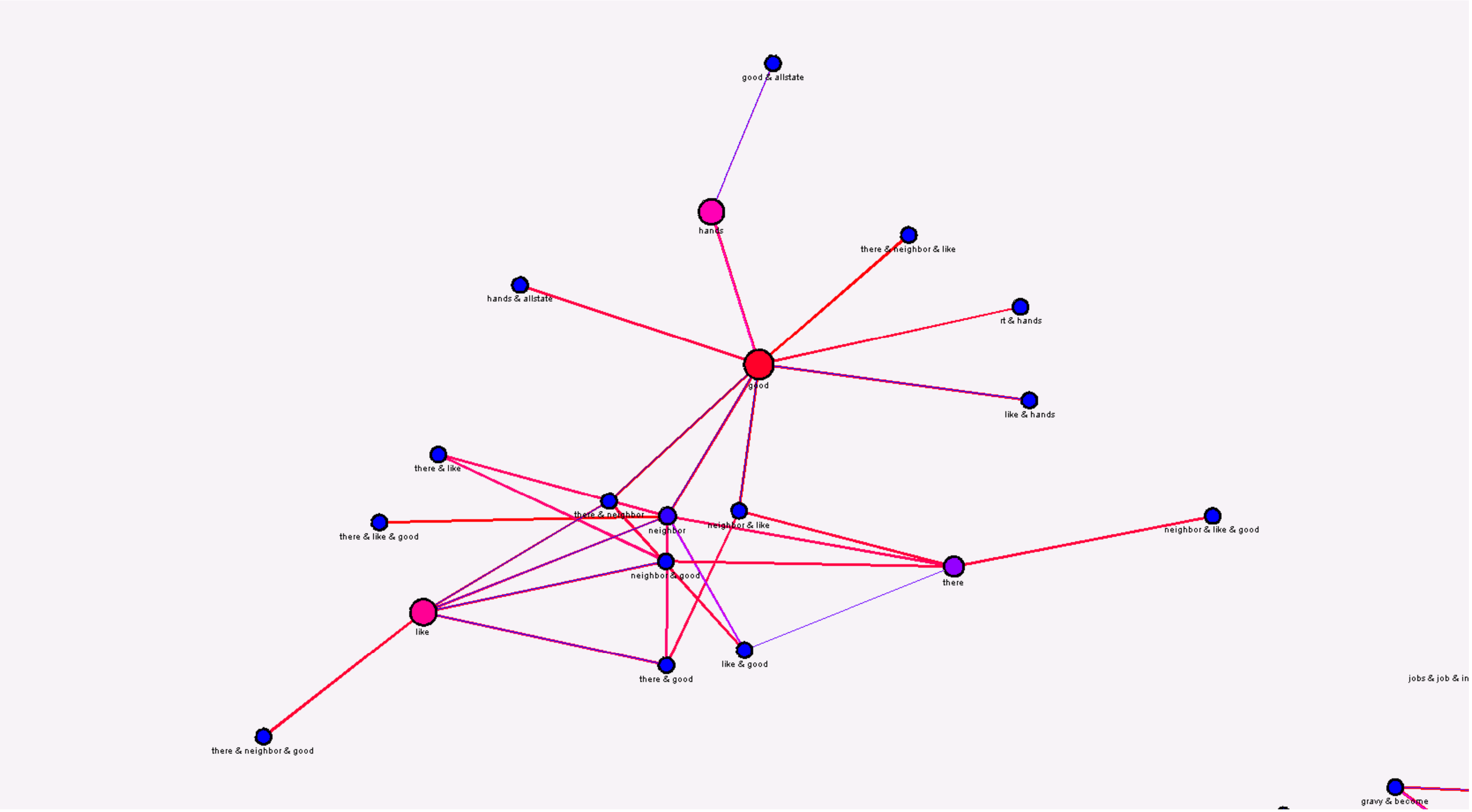
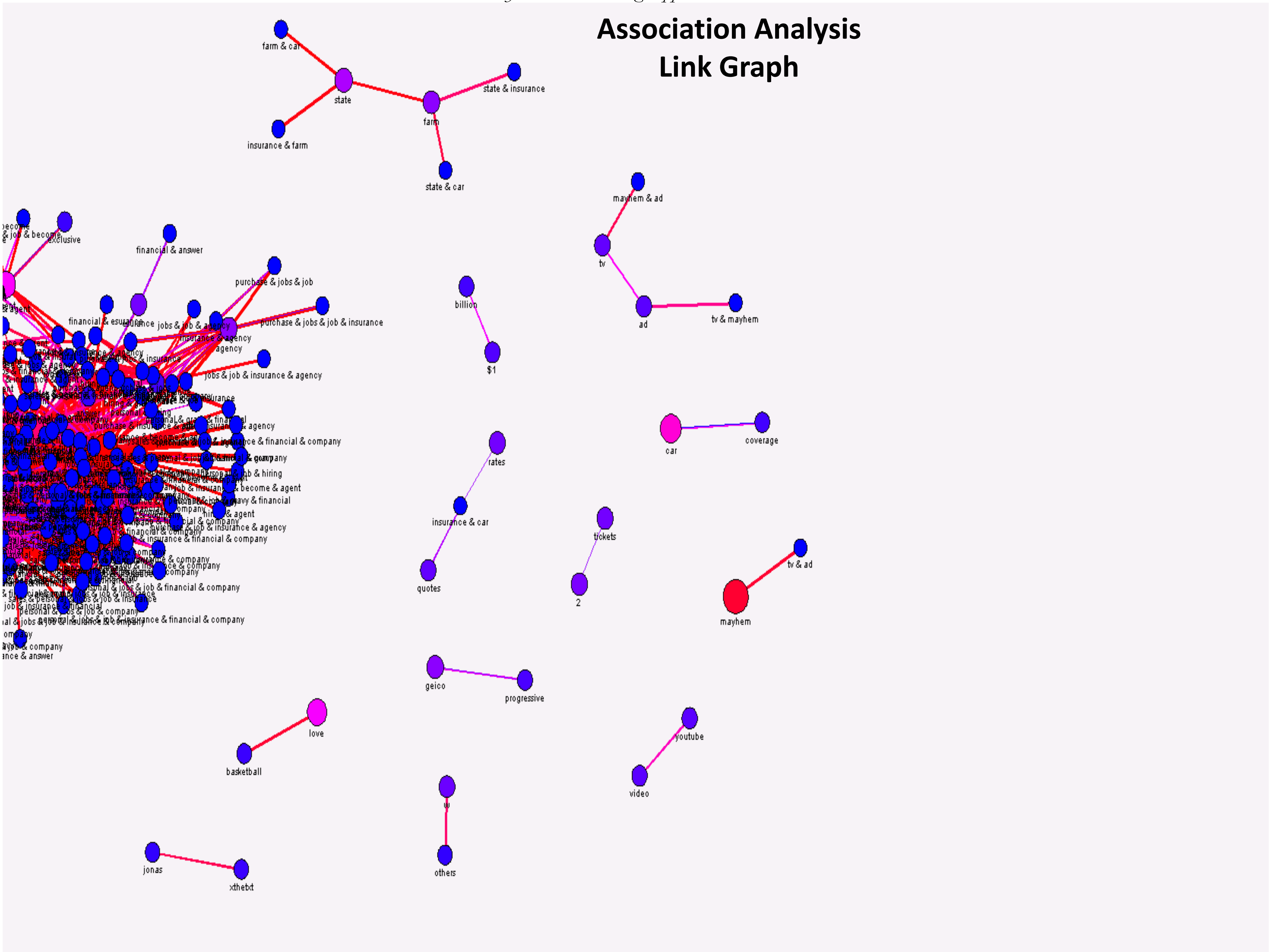


Exhibit 7

Association Analysis Link Graph



Tweet Locations

Exhibit 8