Paper 306-2012

Analyzing sentiments in Tweets about Wal-Mart's gender discrimination lawsuit verdict using SAS® Text Miner

Hari Hara Sudhan, Satish Garla, Goutam Chakraborty,

Oklahoma State University, OK, USA

ABSTRACT

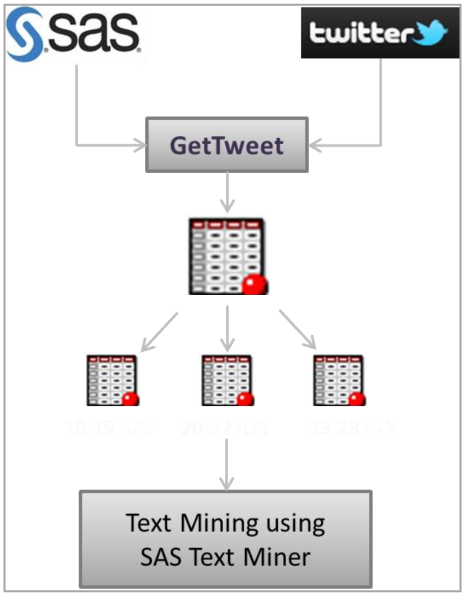
Social Media has gained considerable attention as a valuable source to monitor customers and public reactions following corporate events. Especially the Tweets posted on Twitter are often used to spot trends, moods and sentiments of customers and public. Given the huge volume of tweets that gets posted every day, it is extremely difficult for firms to spot current trends related to public's expressed sentiments about activities of the firm in the tweets. This paper demonstrates the application of a SAS® macro (%GetTweet) to collect and summarize tweets and then an application of sentiment analysis on the fetched tweets using SAS Text Miner using directed search and summarization of specific text items.

INTRODUCTION

Social Media has gained attention in the past decade as a valuable source for information for businesses, governments, and nonprofit organizations across the world. The rate of growth in the number of users of social media sites is in the hundreds of thousands every day. Thus, a treasure-trove of potential information is available from social media if the textual data can be easily accessed and cleaned. The cleaned textual data can then be used for sentiment analysis, cluster analysis, classification analysis and concept trending. Wal-Mart Inc. had recently won a gender discrimination lawsuit. There were immediate and mixed (positive and negative) responses among the public about the verdict which was evident in their tweets. By analyzing this data (tweets) we could explore the sentiments of the public about Wal-Mart before and after the verdict was passed. Text mining was performed on the data that was received before the day of verdict, on the day of verdict and couple of days after the verdict was passed. Clustering the tweets in each time period, along with concept link diagram, we could spot the differences in sentiments expressed by the people on Twitter.

DATA

The %GetTweet macro uses the HTTP procedure in SAS® to communicate with Twitter via the Search API provided by Twitter. The tweets that match the search criteria of the macro are fetched in XML format which is then converted as SAS data set. We used this macro to collect tweets posted about Wal-Mart during 18 JUN 2011 to 28 JUN 2011.As mentioned in Figure 1 the data was split into three groups: tweets posted before 20 JUN 2011(GROUP 1), tweets posted between 20-22 JUN 2011(GROUP 2) and tweets posted between 22-28 JUN 2011 (GROUP 3). A paper was published in the SAS Global forum 2011 where a macro was created to get tweets from Twitter website [1]. This macro was used to get the tweets on Wal-Mart. The tweets were collected and then stored as three datasets (identified as three groups in the schematic diagram below).



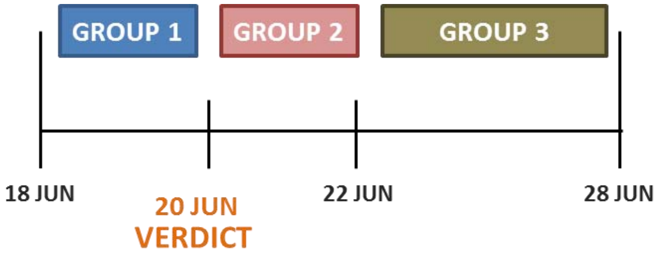


Figure 1 Grouping of data

Text Mining was then performed on each group of data separately. Clusters were formed in each time period using SAS® Text Miner and terms relating to clusters were identified. Based on the concept link diagram the terms associated with the clusters helped us identify the expressed sentiment of the people. Data cleaning of the tweets was done using PERL Regular Expressions. The %GetTweet macro generates a two page summary report, but does not produce the clustering that is needed to extract sentiments out of the tweets.

TEXT MINING

SAS® Text Miner supports various sources of data such as Textual files, SAS Datasets that has text as observations and external databases [3].Text Mining starts with text parsing which identifies unique terms in the text variable and identifies parts of speech, entities, synonyms and punctuation. The Text Miner node performs the parsing and analyzing text data and thereby prepares the data for predictive modeling or further exploration. A typical text mining problem has more terms than documents resulting in a sparse rectangular terms-by-document matrices [3].
A stop list is a collection of words that you want to remove from the text, which has been saved as a SAS dataset. SAS allows you to customize the existing stop lists. Even after using customized stop lists in a corpus of several thousands of documents, the term-by-document matrix can contain hundreds and thousands of terms. It becomes computationally very difficult to analyze a matrix with high dimensional sparse data. Singular Value Decomposition (SVD) can be used to reduce the dimensionality by transforming the matrix into a lower dimensional and more compact form [3]. As a general rule, smaller values of k (2 to 50) are useful for clustering, and larger values (30 to 200) are useful for prediction or classification [4]. For this study, the SVD dimensions were specified as 50.

Using these SVD values the clusters are formed such that the documents in a cluster are more similar to each othere and documents across clusters are less similar to each other. The terms along with their weights are used for creating these groups. Each group or cluster is represented by a list of terms, and those terms will appear in most of the documents within the group [3]. SAS® Text Miner uses Expectation-Maximization algorithm for clustering.

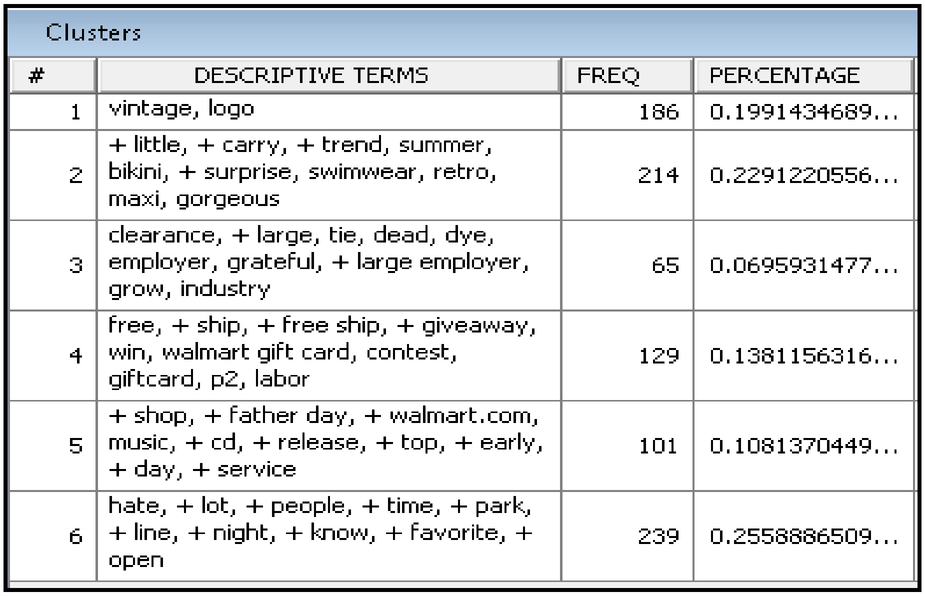
RESULTS

The clusters formed in each group of data were used for comparing the results. For each group of data the number of clusters was specified as 6 (this required some trial-and-error to get meaningful cluster). The SVD values were used to form the clusters. A synonym list was initially drafted based on the first group. Based on the terms that were identified in the second and third group, the synonym list was modified. A single synonym list was used for all the groups. In the case of stop lists, custom stop lists were used for each group. Terms that were less frequent were removed from forming clusters. In such cases each group had their own set of terms that had to be removed from forming clusters and hence we used separate stop lists for each group. The descriptive terms identified in each terms were used for understanding the clusters. The clusters formed in each group are described below.

GROUP 1: TWEETS COLLECTED BEFORE THE VERDICT

The tweets that were collected before the verdict was passed were analyzed using the SAS® Text Miner. The descriptive terms for each cluster as identified by SAS Text Miner, which appear in Figure 2, were used to summarize the tweets as follows:

- Positive comments about the products available in Wal-Mart
- Negative comments about the services/facilities offered at Wal-Mart
- Different coupons offered by Wal-Mart are discussed and shared
- Few good reviews and some complaints about products available on Father's day
- Complaints about long checkout lines and people hating to go to Wal-Mart at night



Clusters

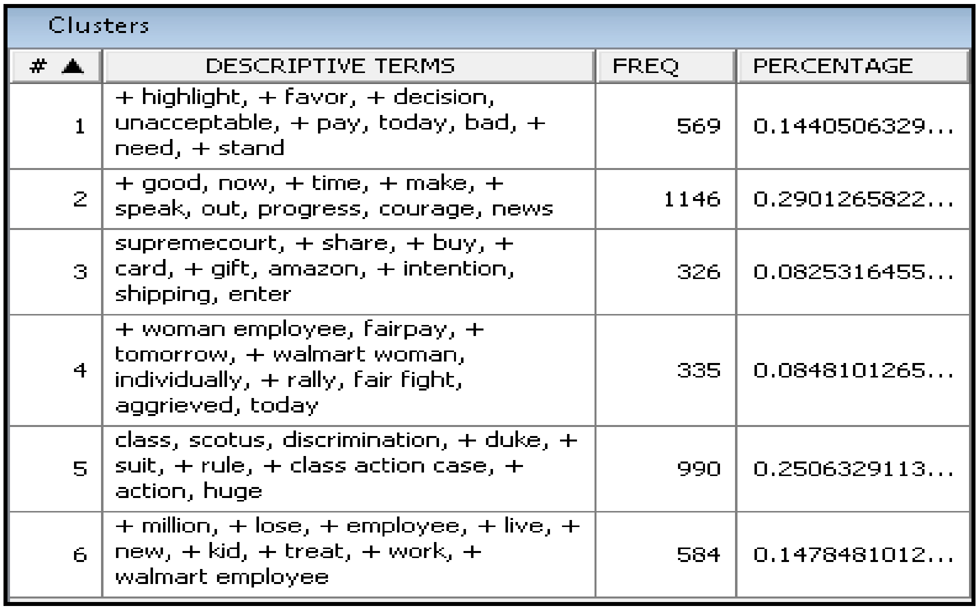
#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE
1	vintage, logo	186	0.1991434689...
2	+ little, + carry, + trend, summer, bikini, + surprise, swimwear, retro, maxi, gorgeous	214	0.2291220556...
3	clearance, + large, tie, dead, dye, employer, grateful, + large employer, grow, industry	65	0.0695931477...
4	free, + ship, + free ship, + giveaway, win, walmart gift card, contest, giftcard, p2, labor	129	0.1381156316...
5	+ shop, + father day, + walmart.com, music, + cd, + release, + top, + early, + day, + service	101	0.1081370449...
6	hate, + lot, + people, + time, + park, + line, + night, + know, + favorite, + open	239	0.2558886509...

Figure 2 Clusters formed in GROUP 1

GROUP 2: TWEETS COLLECTED ON THE DAY OF VERDICT

The tweets that were collected on the day of the verdict were analyzed using the Text Miner. Again, the descriptive terms for each cluster available in Figure 3, as identified by SAS® Text Miner can be used to summarize the tweets as follows:.

- Negative comments about the wages given by Wal-Mart
- Negative comments about the Supreme Court's verdict and many users encouraging women to speak out
- Discussion about the Supreme Court's decision on gender discrimination case and also about the coupons offered by Wal-Mart.
- Negative comments about the verdict. Even after the verdict, rallies were conducted across US to protest.
- People announce the Supreme Court verdict that went in favor of Wal-Mart in their tweets. No specific tone presented.
- Comments about the 1.5 million women whose lawsuit was denied. Also, negative comments about the low percentage of women who work as managers at Wal-Mart.



Clusters

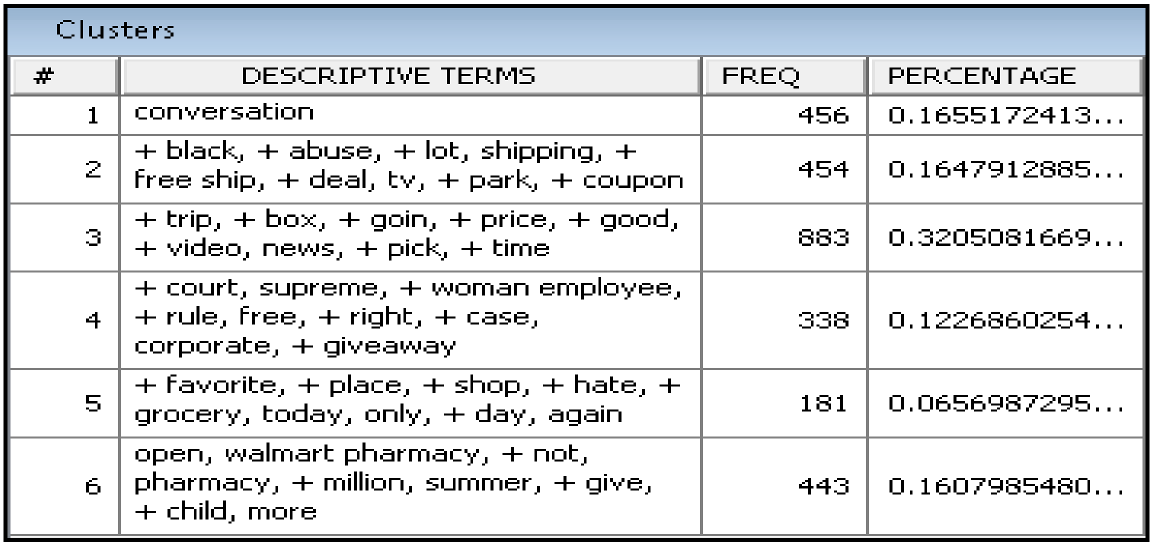
# ▲	DESCRIPTIVE TERMS	FREQ	PERCENTAGE
1	+ highlight, + favor, + decision, unacceptable, + pay, today, bad, + need, + stand	569	0.1440506329...
2	+ good, now, + time, + make, + speak, out, progress, courage, news	1146	0.2901265822...
3	supremecourt, + share, + buy, + card, + gift, amazon, + intention, shipping, enter	326	0.0825316455...
4	+ woman employee, fairpay, + tomorrow, + walmart woman, individually, + rally, fair fight, aggrieved, today	335	0.0848101265...
5	class, scotus, discrimination, + duke, + suit, + rule, + class action case, + action, huge	990	0.2506329113...
6	+ million, + lose, + employee, + live, + new, + kid, + treat, + work, + walmart employee	584	0.1478481012...

Figure 3 Clusters formed in GROUP 2

GROUP 3: TWEETS COLLECTED AFTER THE VERDICT WAS ANNUOUNCED

The tweets that were collected after the verdict was passed were analyzed using the SAS Text Miner. Based on the clusters formed as shown in **Figure 3** the following conclusions were made on the clusters.

- Negative comments about court ruling in favor of Wal-Mart
- Discussion about their shopping experiences at Wal-Mart
- General comments about Wal-Mart such as sale, coupons and experiences at Wal-Mart.
- Discussion about donations that was given by Wal-Mart for child education.



Clusters

#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE
1	conversation	456	0.1655172413...
2	+ black, + abuse, + lot, shipping, + free ship, + deal, tv, + park, + coupon	454	0.1647912885...
3	+ trip, + box, + goin, + price, + good, + video, news, + pick, + time	883	0.3205081669...
4	+ court, supreme, + woman employee, + rule, free, + right, + case, corporate, + giveaway	338	0.1226860254...
5	+ favorite, + place, + shop, + hate, + grocery, today, only, + day, again	181	0.0656987295...
6	open, walmart pharmacy, + not, pharmacy, + million, summer, + give, + child, more	443	0.1607985480...

Figure 3 Clusters formed in Group 3

The clusters formed in each group give us a broad-based idea of what the people tweeted about and if an underlying theme or pattern can be deciphered. For example, tweets in GROUP 1 cluster are mainly about the shopping experience, coupons and good reviews about Wal-Mart. Tweets in GROUP 2 clusters are about the Supreme Court's verdict and the low wage given to female employees. While tweets in GROUP 3 clusters are mainly about the shopping experience in Wal-Mart and a few about the Supreme Court verdict. Although analyzing the clusters gives us only a broad view of the comments, analyzing the frequency of some terms in each group often give us more additional insights about these tweets.

Exploring the frequency of occurrence of important terms in a line plot helps in understanding the trend for the terms across the three data groups as in **Figure 4**. The terms **'court'** or **'supreme'** had been quoted most in GROUP 2. The term **'abuse'** (this is a replacement word we are using for the unprintable bad words in tweets) has been used the most in GROUP 3 and the term **'coupon'** has been used at more or less similar frequency for all groups. We can explore the trends in relationship between important terms across three data groups using concept links as shown in Appendix.

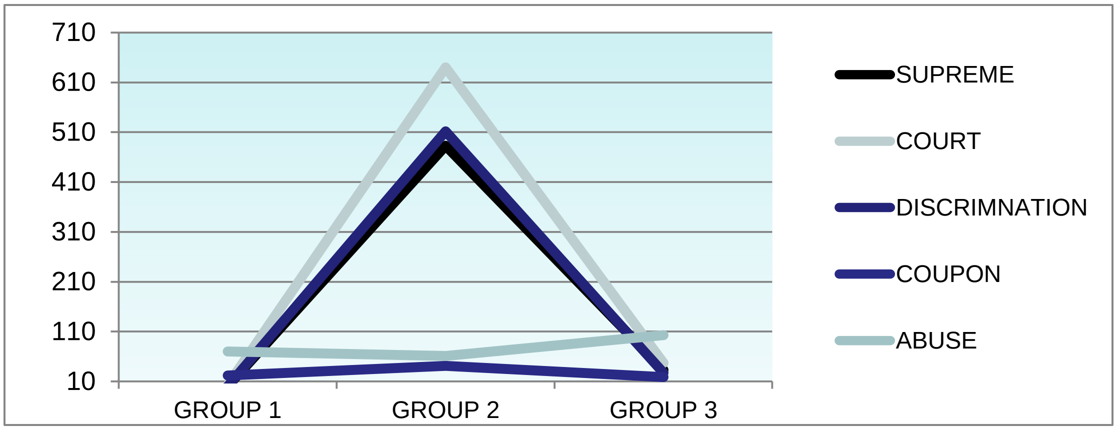


Figure 4 Frequency of terms in each GROUP

CONCEPT LINK DIAGRAM

The concept link diagrams for the three sets of tweets are shown below. The term **walmart** has been chosen for all the datasets (group 1 2 and 3) to illustrate the associations between this term and others terms in those data sets. The thickness of the link between two terms show the strength of association between the two terms. That is, a thick link between two terms show a stronger association than a thin link. Based on the diagram we can see that the term walmart was not associated with **court** or **case** in the first dataset. But, the second and third dataset the same term is associated with Supreme Court, discrimination etc...

The term *walmart* is associated with terms like *abuse* and *buy* in the GROUP 1 dataset as seen in **Figure 5**. In GROUP 2 the term *walmart* is associated with terms like *Supreme Court* and *discrimination* as seen in **Figure 6**. In GROUP 3 *walmart* is associated with terms such as *court, supreme and buy* as shown in **Figure 7**.

GROUP 1

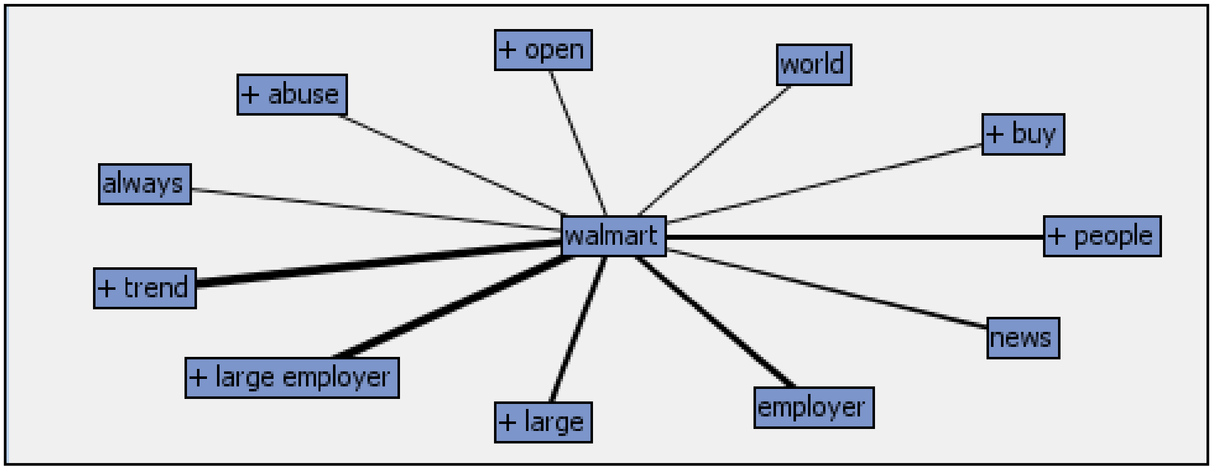


Figure 5: Concept link diagram for GROUP 1

GROUP 2

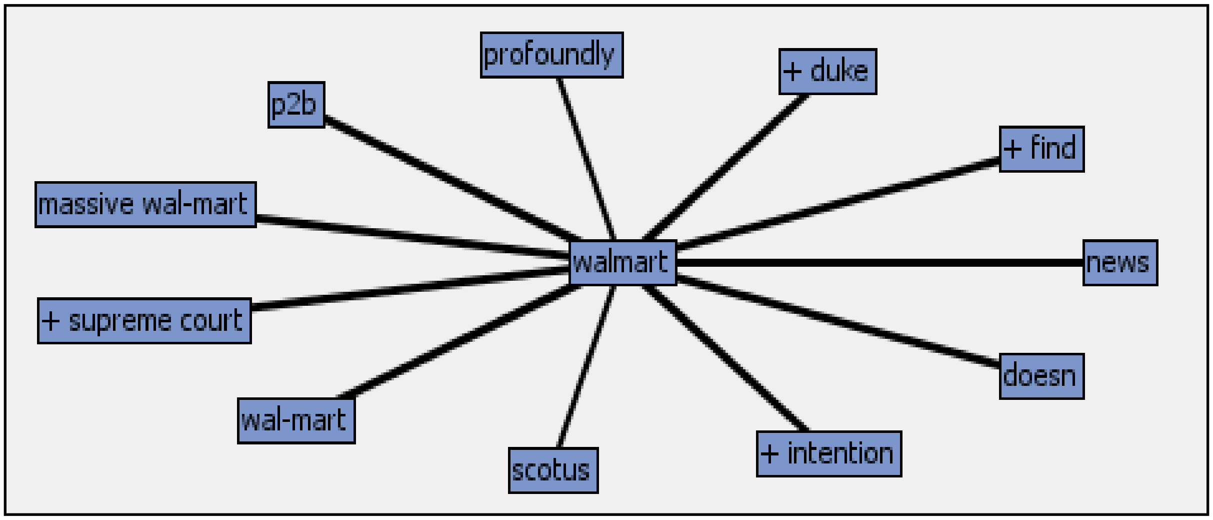


Figure 6: Concept link diagram for GROUP 2

GROUP 3

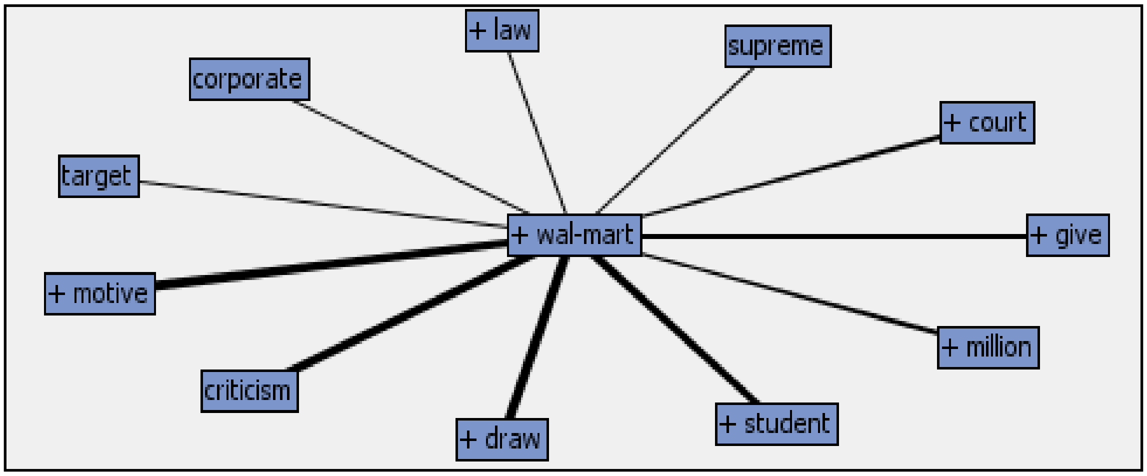


Figure 7: Concept link diagram for GROUP 3

CONCLUSION

Tweets are a great source for exploring the sentiments and trends expressed about any firm. Especially when any important event (external or internal) related to a firm takes place, one can use tweets to spot the public's reactions to such events. In our analysis, we clearly observed the change in topics and sentiments expressed in such topics shared on Twitter before and after the Supreme Court's verdict about Wal-Mart. Text Miner can be used to a certain extent for identifying the sentiments in the text as we have done in this paper. However, our approach involves manual examination of the terms in the clusters which is at a high level of aggregation. Sophisticated sentiment analysis software (such as the one available as a solution package from SAS®) using a dedicated content management system that contains predefined positive and negative terms can surely reveal better insights. As academics, we did not have access to such SAS software at the time of this research and hence we used directed search of specific terms to identify sentiments.

Social Media comes with lot of challenges. In particular, tweets include a lot of chat slangs, URLs and shorthand due to the 140 character limitation on the tweet. Although we used regular pearl expression to clean the tweets, a specialized algorithm to clean social media data would have been helpful for a richer and deeper analysis.

REFERENCES:

[1] Satish Garla, Goutam Chakraborty.2011." %GetTweet: A New SAS® Macro to Fetch and Summarize Tweet." SAS Global Forum 2011, Las Vegas, NV

[2] Ronen Feldman, James Sanger.2007. "Advanced approaches in analyzing unstructured data"

[3] "Introduction to Text Miner" In SAS Enterprise Miner Help.SAS Enterprise Miner 6.2. SAS Institute Inc., Cary, NC.

[4] Annette Sanders, Craig DeVault. 2004. "Using SAS® at SAS: The Mining of SAS Technical Support". Cary, NC: SAS Institute Inc.

[5] Battioui, C. 2008. "A Text Miner analysis to compare internet and medline information about allergy medications. SAS Regional Conference".

APPENDIX

%GETTWEET MACRO

This macro allows anyone to collect customized data from Twitter based on combinations of options as shown below. All the Tweets in the last week that match the search conditions are downloaded into a SAS data set. These conditions are specified as the keyword parameters in the Macro. The full Macro code is reported in the Appendix.

For this study, tweets related to walmart were gathered. Similarly the dates based on each group were specified in the above mentioned macro.

The GetTweet macro function is as follows:

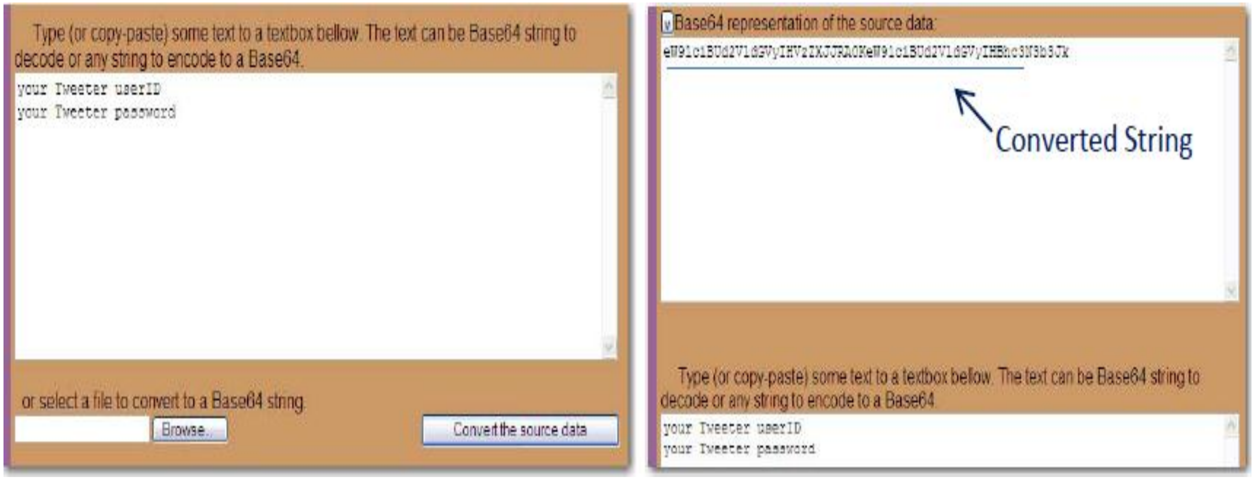
%GetTweet(WORDS=,PHRASE=,ANY=,NONE=,HASH=,FROM=,TO=,SINCE=,UNTIL=,QUESTION=, CODE=,PATH=);

Where

WORDS= (Mention all of the words you want to search. This is an AND condition)
SINCE= (Enter From Date in the format: YYYY-MM-DD)
UNTIL= (Enter To Date in the format: YYYY-MM-DD)
PHRASE= (Enter Exact Phrase you want to search)
ANY= (Enter Any of the words you want to search)
NONE= (Enter the words you do not want to be in search results)
HASH= (Enter the hash tag that you want in your results)
FROM= (Enter name of the person who is Tweeting. Multiple names not allowed)
QUESTION= (Enter 1 if you want only the Tweets with a Question Mark)
CODE= (Enter base64 encoded string of Twitter login – explained below)
PATH= (Directory where fetched data sets will be saved)

In general, Twitter returns up to a week of Tweets with a maximum of about 1,500 Tweets. If search results exceed 1,500 Tweets, only the most recent 1,500 will be kept. To collect Tweets for a complete week on terms that may exceed the 1,500 Tweets limit, one can use the SINCE and UNTIL parameters.

Many websites, including Twitter, require basic authentication to access their database. While there are multiple ways to do this, the macro discussed in this paper uses a base64 encoded Twitter username and password that is passed on as a part of the HTTP header in the PROC HTTP procedure. The base64 encoded string is usually in the form "dXhsdgfhsd….". To find the base64 encoded string for your own Twitter login and password, go to the website: http://www.motobit.com/util/ base64-decoder-encoder.asp , type your Twitter ID and Password in the box, and click "convert the source data" button as shown below.
Converted String



Type (or copy-paste) some text to a textbox bellow. The text can be Base64 string to decode or any string to encode to a Base64.

```
your Tweeter userID
your Tweeter password
```

or select a file to convert to a Base64 string

Base64 representation of the source data:

```
eW91ciBUd2VldGVyIHVzZXJJRA0KeW91ciBUd2VldGVyIHBhc3N3b3Jk
```

Converted String

Type (or copy-paste) some text to a textbox bellow. The text can be Base64 string to decode or any string to encode to a Base64.

```
your Tweeter userID
your Tweeter password
```

To use the macro, you need to enter the converted string as one of the parameters.

CONTACT

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Hari Hara Sudhan
Enterprise: Oklahoma State University
Address: Department of MSIS, Oklahoma State University
City, State ZIP: Stillwater, OK - 74074
Work Phone: (720)351-2518
E-mail: hari.duraidhayalu@okstate.edu

Hari Hara Sudhan is a Master's student in Management Information Systems at Oklahoma State University with specialization in Data Mining and Business Intelligence Tools. Has three years of professional experience as System Engineer at Infosys Technologies Ltd.

Name: Satish Garla
Enterprise: Oklahoma State University
Address: Department of MSIS, Oklahoma State University
City, State ZIP: Stillwater, OK - 74074
Email: satish.garla@okstate.edu

Satish Garla is working as Risk Consultant for SAS Institute. He has three years of professional experience as Oracle CRM Consultant. He is SAS® Certified Advanced Programmer for SAS® 9 and Certified Predictive Modeler using SAS® Enterprise Miner 6.1.

Name: Dr. Goutam Chakraborty
Enterprise: Oklahoma State University
Address: Department of Marketing, Oklahoma State University
City, State ZIP: Stillwater, OK - 74074
Work Phone: (405)744-7644
E-mail: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as *Journal of Interactive Marketing*, *Journal of Advertising Research*, *Journal of Advertising*, *Journal of Business Research*, etc. He chaired the national conference for direct marketing educators in 2004 and 2005 and co-chaired the M2007 data mining conference. He is also a Business Knowledge Series instructor for SAS.

TRADEMARKS