Paper 1775-2014

Application of Text Mining in Tweets Using SAS®and R, and Analysis of Change in Sentiments toward Xbox One Using SAS®Sentiment Analysis Studio

Aditya Datla, Oklahoma State University; Reshma Jayarajan Palangat, Oklahoma State University; Dr. Goutam Chakraborty, Oklahoma State University;

ABSTRACT

The power of social media has increased to such an extent that those businesses that fail to monitor consumer responses on social networking sites are now clearly at a disadvantage. In this paper, we aim to provide some insights on the impact of the Digital Rights Management (DRM) policies of Microsoft and the release of Xbox One on their customers' reactions. We have conducted preliminary research to compare the basic text mining capabilities of SAS® and R, two very diverse yet powerful tools. Our results suggest that SAS® works better than R when it comes to extensive analysis of textual data. Furthermore, we performed sentiment analysis on Tweets collected before and after the launch of Xbox One and found a positive valence in the opinions. Probing further into the sentiments (using Feature based analysis) of customers pertaining to the controller, we see a negative response perhaps attributed to the comfort level in comparison with the PS4 controller.

INTRODUCTION

Our analysis is split into two main segments. For the first part, a total of 6,500 Tweets were collected to analyze the impact of the DRM policies of Microsoft. The Tweets were segmented into three groups based on the date: before Microsoft announced its Xbox One policies (May 18 to May 26), after the policies were announced (May 27 to June 16), and after changes were made to the announced policies (June 16 to July 1). Once the analysis was performed on these Tweets, the result obtained from both SAS® and R was compared.

As for the second part, to further expand the scope of this project, we collected Tweets from before and after the announcement to launch the Xbox was made to get a sense of the sentiments involved. Because R does not have the extensive ability to perform sentiment analysis, we decided to use only SAS® Sentiment Analysis Studio to serve this purpose. Tweets posted before and after the release of Xbox One were collected, resulting in two categories of Tweets; specifically those posted between November 15 and November 21 followed by those posted between November 22 and November 29.

PART 1: COMPARING TEXT MINING CAPABILITIES OF SAS® AND R

DATA EXTRACTION:

Using web scraping techniques, we collected 6,500 Tweets, which were then segmented into three groups based on the dates on which they were tweeted, as shown below:

• Before Microsoft announced its Xbox one policies [1,686 Tweets collected between May 18th and May 26th]

• After the policies were announced [2,946 Tweets collected between May 27th and June 16th]

• After changes were made to the announced policies [1,868 Tweets collected between June 16th and July 1st]

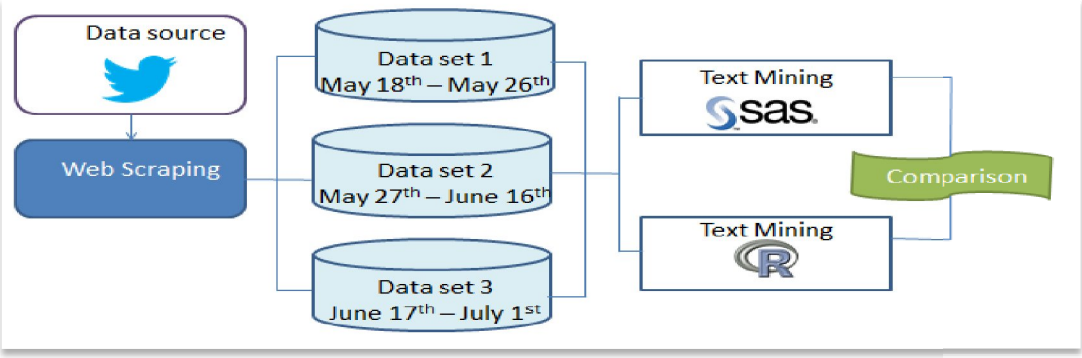


Figure 1: Text mining Frame Work

DATA PROCESSING USING SAS® EM:

Initially, the data set containing Tweets was parsed using text parsing nodes in order to quantify the words used in the Tweets, analyze and filter out stop words, special characters, etc. After determining stemming words, a custom list of synonyms was created. Next, a text filtering node was used to achieve the default frequency weighting using default term weights. Following this, the default dictionary, with custom words added to it, was used. This node reduces the list of parsed terms to the most relevant terms. To find the various topics, the output obtained from the text filter node was then passed through the text topic node, where a trial and error method was employed to reduce multi term topics to five.

DATA PROCESSING USING R:

In order to perform text mining in R, packages such as Twitter, tm (text mining), wordcloud, & RColorBrewer were downloaded and added onto the tool. The unstructured data was processed by converting it to plain text documents, while eliminating white spaces, stopwords, and stemming and filtering custom words. After this a Term Document Matrix (TDM) was created from the corpus. It contains the list of all the words left from the filtering process along with their frequency of occurrence in the TDM. Using the findFreqTerms() function, the most frequent terms were identified by mentioning a cutoff value. Wordcloud was used to visually notice the focus on the most regularly quoted words in a text. Relevant words were chosen based on the wordcloud obtained. Subsequently, the function findAssocs() was used to get the most associated list of words for the selected frequent words.

ANALYSIS

The topics generated in data set 2, after the policy was announced, have a negative connotation as compared to those in data set 1 and 3, i.e., before and after the changes were made to policy, implying that Microsoft's decision to reverse the policy did, in fact, have an impact on customer views. After Microsoft announced the DRM policy, PS4 became the highest pre-ordered console on Amazon, but after Microsoft reversed the policy, Xbox regained its top position within 24 hours. By monitoring customer views on social media, Microsoft was able to quickly recoup from a potential loss in revenue and a damaged reputation.

Topic ID	Topic
1	+microsoft,+unveil,+windows,tv,+knov
2	+game,xbox,+play,+want,+good
3	+day,+video,sony,+release,youtube
4	+check,+game,+release,+console,+spot
5	+achievement,+xbox achievement,+unlock,+dead,walking

Topic ID	Topic
1	Xbox Policy, privacy, ban, failure, restriction, hate
2	+playstation,+analyst,cost,ps4,+price
3	+win,+detail,+follow,console,+contest
4	xbox,+xboxsupport,+buy,ps4,+ban
5	ps4,+e3,+game,+sony,+day
6	+microsoft,+reddit,+positive,+comme

Topic ID	Topic
1	+follow,gamerdeals,+win,console,+giveaway
2	drm,+policy,+microsoft,+reverse,+restriction
3	kinect,+long,+require,+function,+microsoft
4	+ps4,+fav,+good,+controller,fifa
5	+game,live,+gold,+microsoft,+confirm

Figure 2: Topics generated by SAS® EM

COMPARISON RESULTS

Pre-processing

The text pre-processing capabilities of R are on par with those of SAS® EM in terms of eliminating the stopwords, special characters, stemming words & creating a Term Document Matrix. However, SAS® EM offers a range of customization techniques such as detecting/ignoring the parts of speech and creating a custom list of synonyms which in the case of R is tremendously challenging, especially when dealing with large text documents.

Textual Analysis

R provides very limited analysis options compared to what SAS® has to offer. The most frequent terms occurring in the text document can be visualized using both the tools. Concept link diagrams produced by SAS® offer a more interactive way to find the associated terms with a given term and to drill down the hierarchy, unlike R which only gives a list of associated terms with a given term.

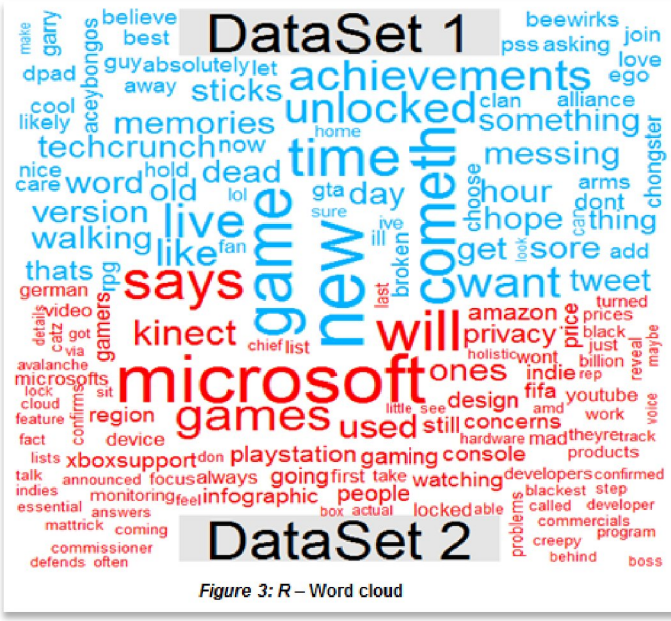


Figure 3: R – Word cloud

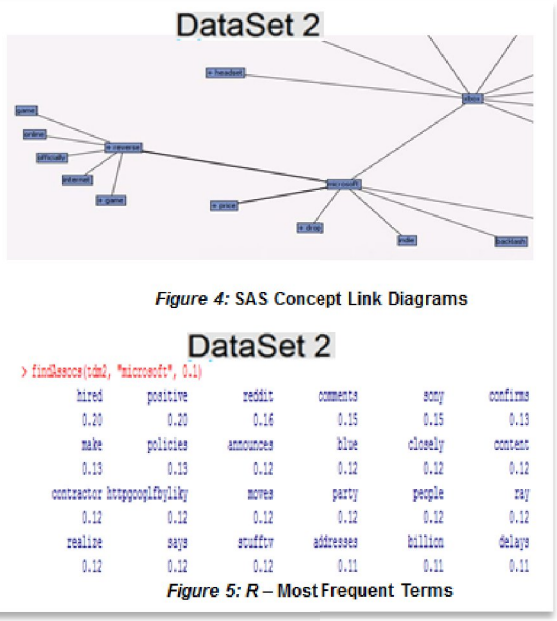


Figure 4: SAS Concept Link Diagrams

Figure 5: R – Most Frequent Terms

SAS® EM produces topics using the text topic node; these topics can be extremely convenient in summarizing the content of the textual data, a functionality that R lacks. Also, SAS® can be used to perform Sentiment Analysis on the data which offers an in-depth analysis of the data. At the time of this research, the "Sentiment" package was removed from the Cran repository in R and hence it was not possible to perform sentiment analysis.

PART 2: SENTIMENT ANALYSIS ON TWEETS ABOUT XBOX ONE USING SAS®

INTRODUCTION

Traversing through the analysis process, we were intrigued by the sentiment analysis capability of SAS®, and hence decided to use a larger dataset to probe deeper. Tweets posted around the time of the launch of Xbox one, ranging from a week prior to and also a week after its launch were collected using FLUME.

DATA EXTRACTION

We collected data from Twitter using FLUME on the Memory Channel (MEM) after which the data is sent to the Hadoop Distributed File System (HDFS) sink, and then to the HDFS as depicted in the architecture diagram below.

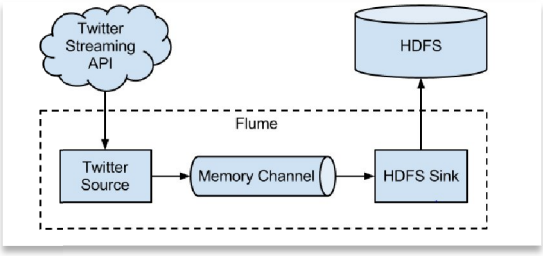


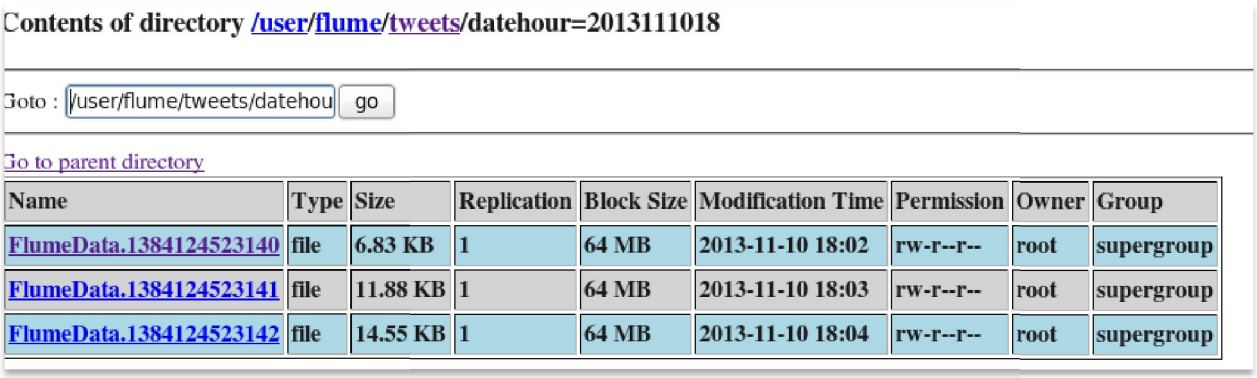
Figure 6: Architecture Diagram

HDFS is used to store data collected from social networking sites. Since the data is in an unstructured format, it is converted into a structured format using Hue and Hive. The required columns can be specified in the Hive queries which will in turn be used for analysis.

We followed the subsequent steps in order to collect the Tweets:

- Created a developer's account on Twitter.
- Used FLUME to collect data from the Twitter API and store it in HDFS.
- Created a table for the Tweetsto store the data into the related columns.
- Created an Oozie workflow to perform the job periodically every 1 hour.

FLUME was used to inject the data into HDFS, which then stores the data in the Data Nodes, Hive application was used to transform the unstructured data into structured data and finally the Map Reduce function was called so as to perform the operation required for analysis. The figure below shows how the data is stored in HDFS.



Contents of directory /user/flume/tweets/datehour=2013111018

Goto : /user/flume/tweets/datehou go

Go to parent directory

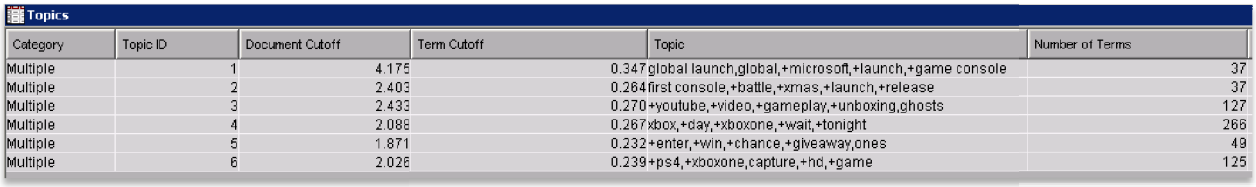
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1384124523140	file	6.83 KB	1	64 MB	2013-11-10 18:02	rw-r--r--	root	supergroup
FlumeData.1384124523141	file	11.88 KB	1	64 MB	2013-11-10 18:03	rw-r--r--	root	supergroup
FlumeData.1384124523142	file	14.55 KB	1	64 MB	2013-11-10 18:04	rw-r--r--	root	supergroup

Figure 7: Data stored in HDFS

DATA PROCESSING

A code was written to exclude all the Tweets that did not include a custom list of words that comprised of conjunctions, prepositions and other frequently used English words. This was done in order to eliminate Tweets that were in languages other than English, especially those which used Latin alphabets such as Spanish, Malay etc.Re-Tweets were excluded from our analysis in order to eliminate any sort of bias in the topics or clusters to be created. Following this, the steps that were followed in PART 1 were carried out in the same manner.

PRE-RELEASE TWEETS:TEXT TOPICS AND TEXT CLUSTERS



Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms
Multiple	1	4.175	0.347	global launch,global,+microsoft,+launch,+game console	37
Multiple	2	2.403	0.264	first console,+battle,+xmas,+launch,+release	37
Multiple	3	2.433	0.270	+youtube,+video,+gameplay,+unboxing,ghosts	127
Multiple	4	2.088	0.267	xbox,+day,+xboxone,+wait,+tonight	266
Multiple	5	1.871	0.232	+enter,+win,+chance,+giveaway,ones	49
Multiple	6	2.026	0.239	+ps4,+xboxone,capture,+hd,+game	125

Figure 8: Text Topics for Pre-release Tweets

Topic ID #1 was primarily based on the global launch of Xbox. Next, Topic ID #2 was about its release and the excitement about receiving an Xbox as a Christmas present. Topic ID #3 was based on YouTube videos perhaps of people unboxing their new Xbox. A vast majority of terms that belonged to Topic ID #4 were generated based on Tweets that were posted on the night before the launch, indicating that people were eagerly waiting for the Xbox and closely monitoring updates about it. Topic ID #5 was centered on giveaways and chances of winning an Xbox. The last topic, Topic ID #6 discussed about HD quality as well as a possible comparison of the Xbox with the PS4.

We decided to use a text clustering node as some Tweets did not belong to any of the topics mentioned above, or fell into multiple topic categories. Using a Single Value Decomposition (SVD) value of 40 we tried to ensure that we did not lose much information and yet reaped the benefits of reduced dimensionality. Other options were left at their default values. In a text cluster node, a particular document is classifiedinto precisely one cluster. From the text clustering node, we were able to form six clusters. Some characteristics of these clusters were as follows:

- One of the clusters focused on some of the gaming experiences of the players on Xbox one.
- Another cluster concentrated mainly on the excitement of receiving an Xbox as a Christmas gift and

how users were monitoring the updates on the release closely.

- Battlefield 4 was another clustering topic, where gamers discussed 'second assault' which is an expansion pack of the existing game.
- Additionally, there was a cluster that was generated based on Tweets revolving around the Xbox One giveaways.

Clusters

Cluster ID ▲	Descriptive Terms	Frequency	Percentage
1	'dead end' +battlefield +dead +deo +duty +end +forza +gameplay +giveaway +graphics +kinect +live +multiplayer +rising +rom...	4188	6%
2	+'midnight release' +buy +christmas +fuck +gonna +know +line +lol +money +people +play +release +shit +tonight +want ...	12509	19%
3	+christmas +friday +good +jumpahead +kingjames +line +midnight +night +people +play +ready +time +tomorrow +wait +wor...	18404	28%
4	'a lot of' 'pack second assault' +assault +battlefield +community +few +hour +moment +news +pack +second +sur +thursday ...	3425	5%
5	'global launch' +'game console' +console +doctor +early +finally +game +ios +kinect +kingjames +launch +map +microsoft +n...	9383	14%
6	'first console' +battle +chance +console +day +edition +em +enter +game +giveaway +good +hd +live +ps4 +win ...	17627	27%

Figure 9: Text Clusters for Pre-release Tweets

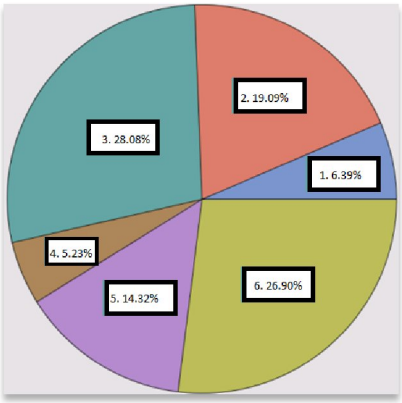


Figure 10: Pie chart of clusters

POST-RELEASE TWEETS:TEXT TOPICS AND TEXT CLUSTERS

Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms
Multiple	1	3.051	0.267	+sell,+million,+hour,+microsoft,+console	138
Multiple	2	3.082	0.270	youtube,+video,+unboxing,+gameplay,xbox one	201
Multiple	3	2.314	0.231	+enter,+win,+ps4,+giveaway,+chance	73
Multiple	4	2.360	0.250	+xbox one,+play,+ps4,+buy,+game	407
Multiple	5	1.986	0.214	+day,+edition,+xbox one,+unboxing,+microsoft	175

Figure 11: Text Topics for Post-release Tweets

Likewise for post-releaseTweets, text topic and text cluster nodes were run. Some of the topics generated were almost similar to the topics generated by the pre-releaseTweets. There were several Tweets about a chance to win giveaways, and this caused numerous re-Tweets, which were filtered out in our analysis. In addition, there were several re-Tweets about Microsoft making sexist remarks in their advertisements for Xbox. Furthermore, an image showing comparison between the graphic quality of Xbox and PS4 went viral on social media. (More details in Appendix).

Six clusters were formed by the text clustering node. Some of the clusters focused on games such as Call of Duty Ghosts, Battlefield 4, FIFA14 etc. Another cluster was about how one million consoles were sold within the first hour of launch. A different cluster focused on the users discussing their alternatives to the Xbox One, such as a PlayStation or an IPhone.

Several of the Tweets were about the functionality of the controller and other factors related to the controller. In two of the six clusters, there was a mention of the word 'controller'. Upon going through a few Tweets, we decided to do a Feature-based analysis because there were mixed reactions from the Tweeters.

Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
1	'xbox one' +'launch day' +amp +battlefield +chance +deo +duty +enter +gameplay +gaming +ghosts +giveaway +gostei +million +p...	38191	27%
2	'xbox 1' +buy +christmas +controller +fuck +gonna +good +know +lol +love +night +people +right +shit +thing ...	45095	32%
3	'xbox one sales' +'first day' +'launch day' +day +edition +first +hit +million +sale +today +tycoon +zoo bensbargains blackfriday days...	7750	5%
4	+ad +comparison +console +drive +help +iphone +join +launch +microsoft +online +party +playstation +review +sony console ...	17919	13%
5	'xbox one' +cable +console +controller +disc +drive +game +gonna +help +install +kinect +online +play +time +tv ...	29050	21%
6	'two_sync next' 'ultimate gaming package' +'ultimate team' +easportsfifa +gaming +gen +legend +netgear +next +night +opening +...	3525	2%

Figure 12: Text Clusters for Post-release Tweets

SENTIMENT ANALYSIS ON POST-RELEASE TWEETS

Upon initial analysis of the Tweets, we figured that although there was a lot of excitement and anticipation with regards to the launch of Xbox One, the pre-releaseTweets failed to showcase any kind of strong emotions. Therefore, we were unable to find sufficient Tweets to classify as negative or positive Tweets. In fact, most of the Tweets had a neutral tendency. In order for us to perform sentiment analysis, it was essential that we had at least some Tweets that had showed a strong manifestation of positivity or negativity. Consequently, we chose not to perform sentiment analysis on the pre-release dataset and thus performed sentiment analysis only on the post-release Tweets.

STATISTICAL MODEL ON POST-TWEETS

In order to build a Statistical Model, we classified about 350 positive and negative Tweets and used these set of Tweets to train the model. The overall distribution of sentiments towards Xbox one was mostly positive. 70.89% of the Tweets were positive and 29.11% were negative.

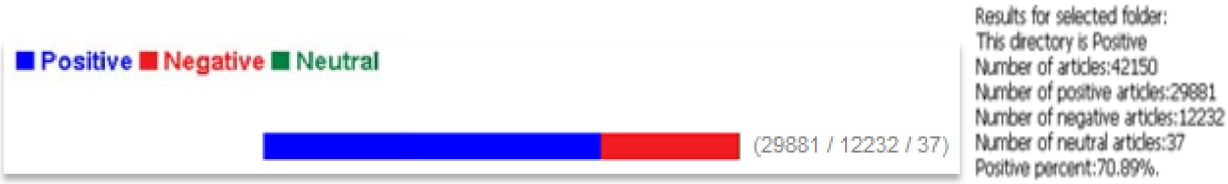


Figure 13: Statistical Model results on Post-release Tweets

SENTIMENT ANALYSIS ON XBOX ONE'S CONTROLLER

Interestingly, we noticed that we could use Feature-based analysis using the term 'controller' in the post-release Tweets,and hence perform sentiment analysis on these Tweets to see if it revealed any distinctive comprehensions.

Since we were only interested in knowing the customers' attitude towards the Xbox One controller, we filtered the Tweets and retained only the ones which had the term 'controller' in them.

- We then went through the 2,600 Tweets which had the term controller in them and classified them as positive and negative Tweets until we had about 100 of each. The neutral Tweets were not classified.
- The positive and negative Tweets were divided into two groups.
- A custom list of positive and negative words was made. These words were scored to represent their degree of positivity and negativity respectively.

After performing the above steps, the custom list of positive and negative words were fed to the Feature-based sentiment analysis model in the SAS®Sentiment Analysis Studio. Also, the folders containing the positive Tweets and negative Tweets were linked to the model. The model was trained using the segmented Tweets to produce the following results for a Statistical Model, Feature-based Model and Hybrid model respectively:

STATISTICAL MODEL ON XBOX ONE'S CONTROLLER

The overall distribution of sentiments towards Xbox one's controller was slightly negative. 46.5% of the Tweets were positive and 53.5% were negative.

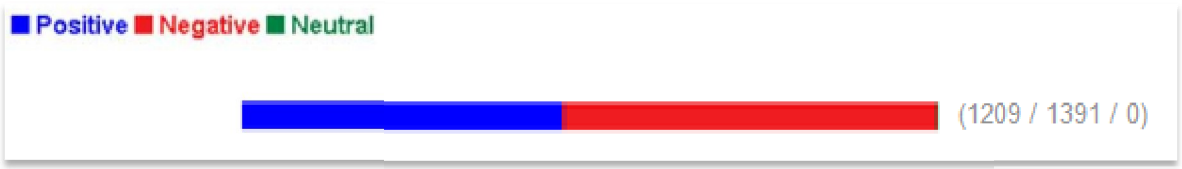


Figure 14: Statistical Model results on Tweets with the word "controller"

RULE BASED MODEL

In order to build a Rule-based model, we imported about 200 negative and 200 positive rules as learned features from the Statistical model that we built on the Xbox One controller's Tweets. Rules that weren't appropriate for our analysis were removed and new rules were added. Some of the excluded rules consisted of numbers and words having neutral sentiments.

In this Rule-based model, we use CLASSIFIER rules along with CONCEPT rules. Since we already mentioned all the positive and negative keywords in the Tonal Keywords section, a simple CONCEPT rule was used to classify Tweets under the feature 'controller'.

The overall distribution of sentiments towards Xbox one's controller was negative. 27% of the Tweets were positive 33% were negative, and 40% were neutral. As for the Feature-based analysis, the results indicated that 37.5% of the Tweets were positive and 62.5% were negative.

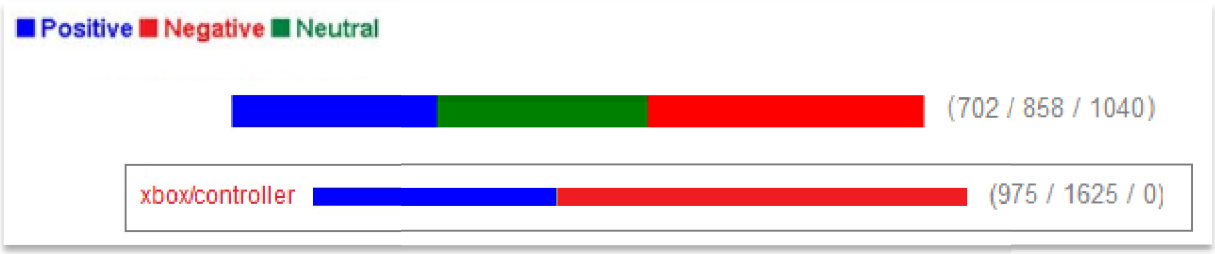


Figure 15: Rule Model results on Tweets with the word "controller"

HYBRID MODEL

The Hybrid model, which is a combination of both Statistical and Rule based models, gave us the following results. The overall distribution of sentiments towards Xbox one's controller was slightly negative. 36.3% of the Tweets were positive and 63.7% were negative.

Figure 16: Statistical Model results on Tweets with the word "controller"

CONCLUSION

To conclude the analysis from Part 1, we believe that the text mining capabilities of SAS are far superior to that of R. The graphics in SAS® are very flexible and easy to use as opposed to the graphics in R which are not attached to statistical analysis and take place separately. Moreover, SAS® has extensive online documentation, professional training, and online support. Packages in R are not written by the development team therefore are not well-refined and could have questionable validity. Sentiment analysis can help companies meet their customers' perceptions of their product by monitoring social media and taking corrective actions if circumstances demand it. As for our

conclusion in Part 2, we see a sense of negativity towards the Xbox One's controller. This information can be taken into consideration to improve future versions of the product in order to meet or exceed customer expectations.

REFERENCES

1. Hari Hara Sudhan, Satish Garla, Goutam Chakraborty. 2012, "Analyzing sentiments in Tweets about Wal-Mart's gender discrimination lawsuit verdict using SAS® Text Miner" SAS Global Forum 2012.

2. Jenn Sykes. 2012, "Predicting Electoral Outcomes with SAS® Sentiment Analysis and SAS® Forecast Studio" SAS Global Forum 2012.

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

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

5. Swati Grover, Jeffin V Jacob, Goutam Chakraborty, 2012, "Analysis of change in sentiments towards Chick-fil-A after Dan Cathy's statement about same sex marriage using SAS® Text Miner and SAS® Sentiment Analysis Studio

6. Aditya Datla, Swati Grover, Goutam Chakraborty, 2012, "Application of Text Mining on Tweets to Analyze the Impact of Xbox One Policies Using SAS and R"

CONTACT

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

Name: Aditya Datla
Enterprise: Oklahoma State University
Address: Spears School of Business, Oklahoma State University
City, State ZIP: Stillwater, OK - 74074
Work Phone: (224) 632-0154
E-mail: adatla@okstate.edu

Aditya Datla is a Master's student in Management Information Systems at Oklahoma State University with specialization in Data Mining. He is a SAS® Certified Base Programmer for SAS® 9, Statistical Business Analyst Using SAS 9: Regression and Modeling and Certified Predictive Modeler using SAS® Enterprise Miner 6.1. In May 2013, he received his SAS® and OSU Business Analytics Certificate.

Name: Reshma Jayarajan Palangat
Enterprise: Oklahoma State University
Address: Spears School of Business, Oklahoma State University
City, State ZIP: Stillwater, OK - 74075
Work Phone: (405) 612-0369
Email: jayaraj@okstate.edu

Reshma Jayarajan Palangat is a Master's student in Business Administration with a specialization in Business Analytics and a Master's student in Telecommunications Management at Oklahoma State University.

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 has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.