

Use of Social Media Data to Predict Retail Sales Performance

Li Zhang, Ph.D., Alliance Data Systems, Inc., Columbus, OH

ABSTRACT

Big data in terms of unstructured social media data is making a big impact in the retail sector. Although recent years have seen increased use of social media data to inform business in the form of qualitative inputs, there is limited work that has spelled out a systematic procedure of using social media data in quantitative analysis. This paper will attempt to fill this gap by showing the practical aspects and theoretical concerns of building analytical models using both structured and unstructured data.

The goals of this paper are modest. First, it provides a brief overview of big data, social media data and sentiment analysis. Second, it offers a classification method that converts and quantifies social media text data into discrete customer sentiment data. Third, this paper describes several statistical models to examine the role of customer sentiment in the prediction of retail sales performance. In addition to the sentiment variables, demographic and econometrics variables are used to account further for variations in retail sales performance. SAS PROC AUTOREG for traditional time series analysis is illustrated throughout this paper. Finally, this paper concludes by highlighting some substantive issues related to using social media data to augment business decisions.

1. INTRODUCTION

Big data has become a popular term in recent years. But what exactly is big data? Troester (2012) briefly defined big data as: "a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making." However, big data is defined more by its variety than its volume. According to a survey over large firms and companies, Davenport and Dyché (2013) emphasized that: "The most important goal and potential reward of big data initiatives is the ability to analyze diverse data sources and new data types, not managing very large data sets." big data resources can include transactions, log data, events, e-mails, social media, sensors, external feeds, RFID scans or POS data, free-form text, geospatial, audio, and still images/videos (Schroek et al., 2012).

Unstructured social media data is a large focus of big data analysis. Arguably, no industry has more to gain from leveraging social media data than the retail sector. Big data technology will help retailers maximize the value of data and gain competitive advantage by listening to the voice of their customers, convert challenges to opportunities, and minimize certain costs and risks. "Customers bring attitudes and expectations shaped by experiences across a diverse commercial world to every interaction with retailers. It is predicted that by 2020, there will be 80 million millennials, as shoppers under 30 years old are known, who have grown up using the Internet for everything" (Mercier et al., 2013).

There is no doubt that the collection of social media data by businesses has become frequent and often necessary. Given this increasing emphasis for social media data, it seems worthwhile to examine some goals and benefits of this seemingly new emphasis. Many marketers and brands have defined these benefits in details. One brief but comprehensive statement of these benefits can be found in an article by Helweh (2011) where the author outlined some key benefits of sentiment analysis of social media data: "Sentiment analysis lets marketers (and market researchers, customer service and support staff, product managers, etc.) get at root causes, at explanations of behaviors that are captured in transactions and tracking records. Sentiment analysis means better targeted marketing, faster detection of opportunities and threats, brand-reputation protection, and the ultimate aim, profit."

It would seem that social media has blown the traditional customer/brand relationship protocol wide open. A fair but obvious question might be "Just how practical are the above statement?" In other words, is there a potential disconnect between these lofty expectations and reality? In the same article, Helweh (2011) highlighted some of the obstacle keeping sentiment analysis from reaching its true potential. The author writes: "Misperceptions, also inflated expectations, fostered by low-grade tools that are keyword based and lack any mechanism to link sentiment to actual business outcomes. On the one hand you get low accuracy, and further there's a 'decision gap.' You get a colorful dashboard, but because the tools are working in isolation, treating social and survey sources as information silos, you can't reliably know what sentiment is important, and what sentiment really means to your business in the sense of driving transactions, boosting satisfaction, and so on."

With all the challenges aside, this paper contributes to the growing body of big data research by showing promising ways that text analytics can be applied to extract and analyze attitudinal information from varied sources, specifically social media data.

In the next section, we present some perspectives on sentiment data analysis to convey the main assumptions and classification procedures used, and the strategy employed to link sentiment data to traditional transaction data. We then describe several statistical models to examine the relationship between sentimentality and retail sales performance. We begin, however, with a broad definition of sentiment data analysis.

2. PERSPECTIVES ON SENTIMENT DATA ANALYSIS

2.1 WHAT IS SENTIMENT DATA ANALYSIS

The primary objective of sentiment data analysis is to replicate the subjective judgments humans make about online, social, and enterprise information sources via the use of natural language processing, text analysis, computational linguistics, or some other modality. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topics or overall polarity of a document. As a point of emphasis, sentiment analysis should be much more than simplistically subtracting the number of “negative” words from the number of “positive” words in a document or message in order to produce a score.

2.2 WHAT ARE THE DIFFERENT SENTIMENT ANALYSIS TECHNIQUES?

The search for the best methods to address sentiment analysis is ongoing and continues to be a major concern for market research and development. Consequently, this increased emphasis has led to the development of several classes of sentiment analysis techniques including keyword spotting, lexical affinity, statistical methods, and concept-level techniques (Cambria et al., 2013).

In this paper, we introduce and expand one class of sentiment data analysis - keyword spotting analysis - using SAS Perl Regular Expressions to locate patterns in text strings and to identify affect. Basic keyword spotting algorithms tend to be linear in their string search strategy and tend to classify text strings by affect categories based solely on the presence of unambiguous affect words, such as like, dislike, and “OMG”. At a basic level, an automated keyword spotting process might involve the following steps:

1. Create a comprehensive list of affective words with corresponding affect categories to be used as reference;
2. Use a text processing tool to pattern-match the reference list to a source data (e.g., string of written text); and finally,
3. Turn the string of written text into simple positive, negative, or neutral sentiments.

2.3 KEYWORD SPOTTING IN SAS ENVIRONMENT

This keyword spotting method could be replicated quite easily in a SAS environment using SAS syntax and Perl Regular Expressions. For example, the text string “I like these shoes!” can be parsed into its individual word components and stored as individual variables to be operated on later. Below is an example of a simplest type of regular expression codes you can use to search for classes of characters (digits, letters, words, etc.) as well as specific character values.

```
DATA _NULL_;
IF _N_ = 1 THEN PATTERN_NUM = PRXPARSE("/\s*like\s*/i");
*match for the word 'like' anywhere in the string;
RETAIN PATTERN_NUM;
INPUT STRING $30.;
POSITION = PRXMATCH(PATTERN_NUM, STRING);
FILE PRINT;
PUT PATTERN_NUM= STRING= POSITION=;
DATALINES;
I like these shoes!
;
RUN;
```

In this simple example, because a match for the word “like” was obtained in the text string by the search algorithm, this string would then be classified as a positive sentiment. The advantage of this linear approach is that it’s very

quick. The disadvantage is that it doesn't consider context when assigning sentiment since it evaluates each word independently of all other words and it doesn't lend itself to comprehensive searching.

Further limiting the utility of single-word search strategies is the fact that sentence structure can be fairly complex and non-monotonic in respect to sentence extension and stop-word substitution (compare "I really like these shoes" vs "I don't really like these shoes" vs "I like these shoes very little").

2.4 AUGMENTED KEYWORD SPOTTING

We address some of the limitations associated with keyword spotting by incorporating a number of rule-based and reasoning-based strategies to the basic search algorithm. These extensions are non-linear in nature and have the added benefit of being able to characterize each affect word based on its relationship with and proximity to its adjacent blocks of texts. This augmented search algorithm will allow for more accurate detection of how often specific words are mentioned and, most importantly, the context in which these words are used.

2.5 EXTENTION OF KEY WORKD SPOTTING WITH CONCEPT-LEVEL TECHNIQUES

Social media data submitted by users can be viewed as a group of words or a single word that attempts to express a complete thought, feeling, or idea. This group of words usually has at least a subject and a main verb. The key concept of sentiment analysis is that some of the word clusters within the text strings are clustered more closely together than other clusters of words in the feeling they attempt to express. In addition, these word clusters tend to associate more strongly or weakly to one of three underlying sentiment dimensions: positive, negative, and neutral sentiments.

There has been increasing interest in the possibility of using concepts derived from structural equation modeling (SEM) to deal with the challenges associated with sentiment data analysis. The SEM techniques provide mathematical and statistical devices that permit researchers to focus on the construct validation of theoretical propositions (Loehlin, 1992).

In our view, the modern SEM techniques can be useful in a number of ways, including

1. Organizing concepts about group of words into scientific models; and
2. Providing tools for the estimation of the components of models.

An initial example of our perspective is illustrated in the table below. Suppose we had to extract meanings from the following text string:

"VERY pleased with the app. Performs much better than most third party apps. Biggest flaw is that the app is HUGE."

Table 1 below is an example of the output of a sentiment SEM analysis with five word clusters and three sentiment factors. The relationship of each word cluster to each underlying factor is expressed by factor loadings.

No.	Word Cluster	Factor 1 (Positive Sentiment)	Factor 2 (Negative Sentiment)	Factor 3 (Neutral Sentiment)
1	<i>"VERY pleased with the app"</i>	0.88	0.13	0.04
2	<i>"Performs much better than"</i>	0.75	0.24	0.2
3	<i>"Biggest flaw"</i>	0.16	0.87	0.06
4	<i>"app is HUGE"</i>	0.28	0.78	0.12
Summary	Average of loadings	0.52	0.51	0.11

Table 1. Example of Sentiment Loadings

The word cluster with the strongest association to the underlying positive sentiment, Factor 1, is *"VERY pleased with the apps"*, with factor loadings of 0.88. Since factor loadings can be interpreted as standardized regression coefficients, one could say that the word cluster *"VERY pleased with the apps"* has a correlation of 0.88 with the Factor 1 (positive sentiment).

The next word clusters, *"Performs much better than"*, also associate strongly with Factor 1 based on their high loadings with this factor. The last two clusters, *"Biggest flaw"* and *"app is HUGE"*, however, have high loadings on Factor 2 (negative sentiment). These two clusters seem to express some specific concerns with the app. Notice that

no cluster value was marginally important in either the positive factor or negative factor and the overall positive and negative factor scores were slightly greater than 0.5. These results suggest that this consumer had a plurality of feelings about this app.

3. SENTIMENT DATA ANALYSIS SYSTEM EVALUATION

Since the primary objective of sentiment data analysis is to replicate the subjective judgments humans make, the accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. Every instance of identifying and extracting subjective information in source materials is a process of reconstruction, and therefore involves some degree of distortion. This process of reconstruction is never random, and is always influenced by the methods or algorithms used to measure and determine the attitudes of the given source material. However, the consistency with which humans produce similar ratings in judging the same abilities or characteristics (called the inter-rater reliability) is typically around 79% (Ogneva, 2010). Therefore, a 70% accurate program would be doing nearly as well as humans.

To evaluate the accuracy of the sentiment analysis, we had $N = 20$ individuals rate 50 randomly selected tweets from the population of tweets collected for this paper. These subjects rated each tweet on two separate dimensions: positive sentiment and negative sentiment. These subjects scored each tweet on both dimensions based on a rating scale of 0 – 3 (0 = Not Applicable, 1 = Slightly, 2 = Moderately, 3 = Strongly). In addition, each tweet is passively scored on a third dimension: neutral sentiment. By default, this neutral sentiment dimension receives a score of zero (0). In the event a subject rated a tweet a score of zero (0) on both the positive sentiment and the negative sentiment dimensions, the neutral sentiment is assigned a value of 1.

4. DATA DESCRIPTION

In this paper, we analyzed data from January 2011 to August 2014 about a retail brand. Plots of all variables data on monthly time series is initially presented in Figures 1 to 8. We have rescaled these data to range from a value of 0 to 1. The plot in Figure 1 shows the data for the outcome variable of retail sales performance (the retail sales on the y-axis against the month on the x-axis). There appears to be a lot of volatility in monthly sales across the 44-month window.

Figures 2 to 4 present the time-series line plots for each of the three types of sentiment variables: positive, negative and neutral sentiments. In each figure of Figures 2 to 4, the upper plots represent the monthly time-series sentiment, and the lower plots represent the same time-series sentiment but include the outcome variable of retail sales performance. We can make three observations about these data. First, the positive sentiment plot shows less volatility than both the negative and neutral sentiment plots. Second, there appears to be some evidence of seasonality. Finally, there appears to be some extreme, though infrequent, sentiment about this brand.

We present a similar type of time-series plots for each of the four additional demographic and economic variables including customer recency, customer age, customer tenure, and treasury 10-year yield rate in Figures 5 to 8 in the appendix section.

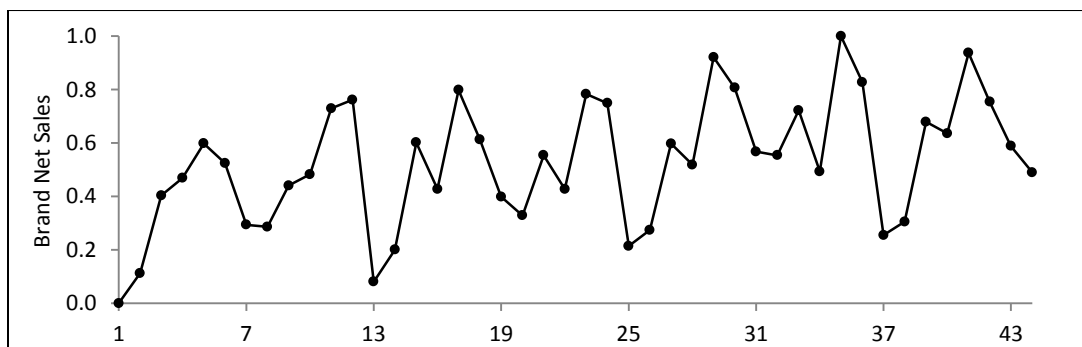


Figure 1. Monthly Time Series of Retail Sales Data

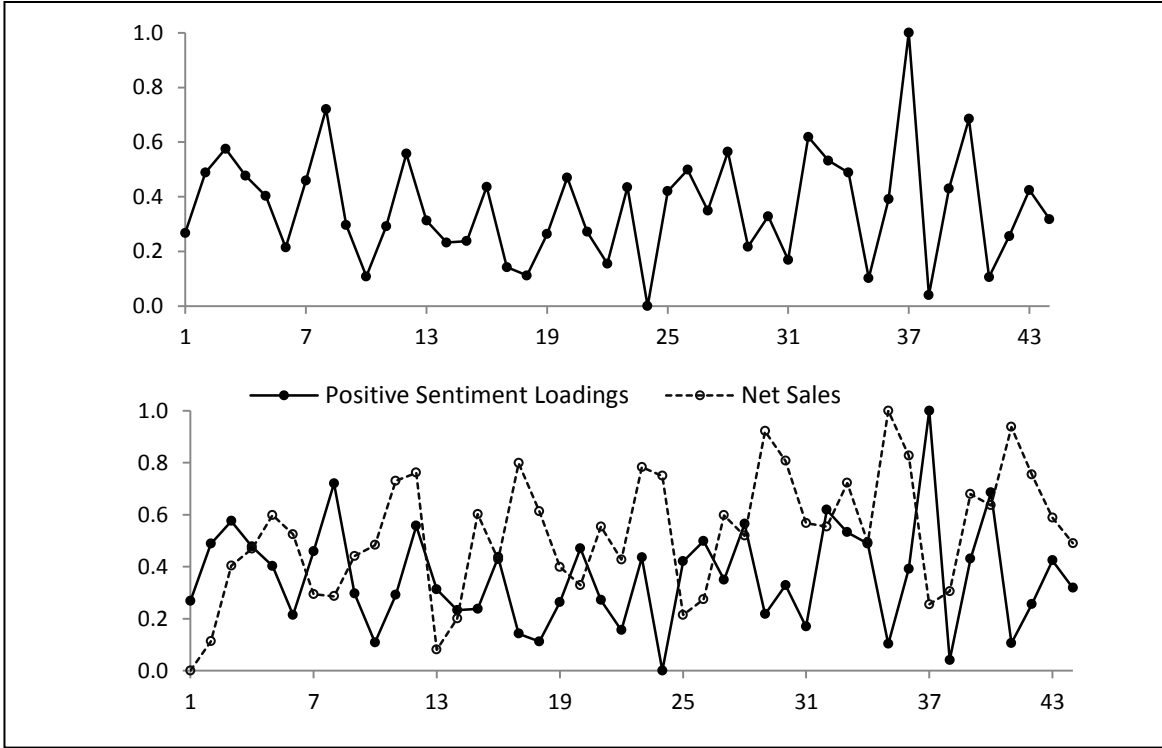


Figure 2. Monthly Time Series of Positive Sentiment Data

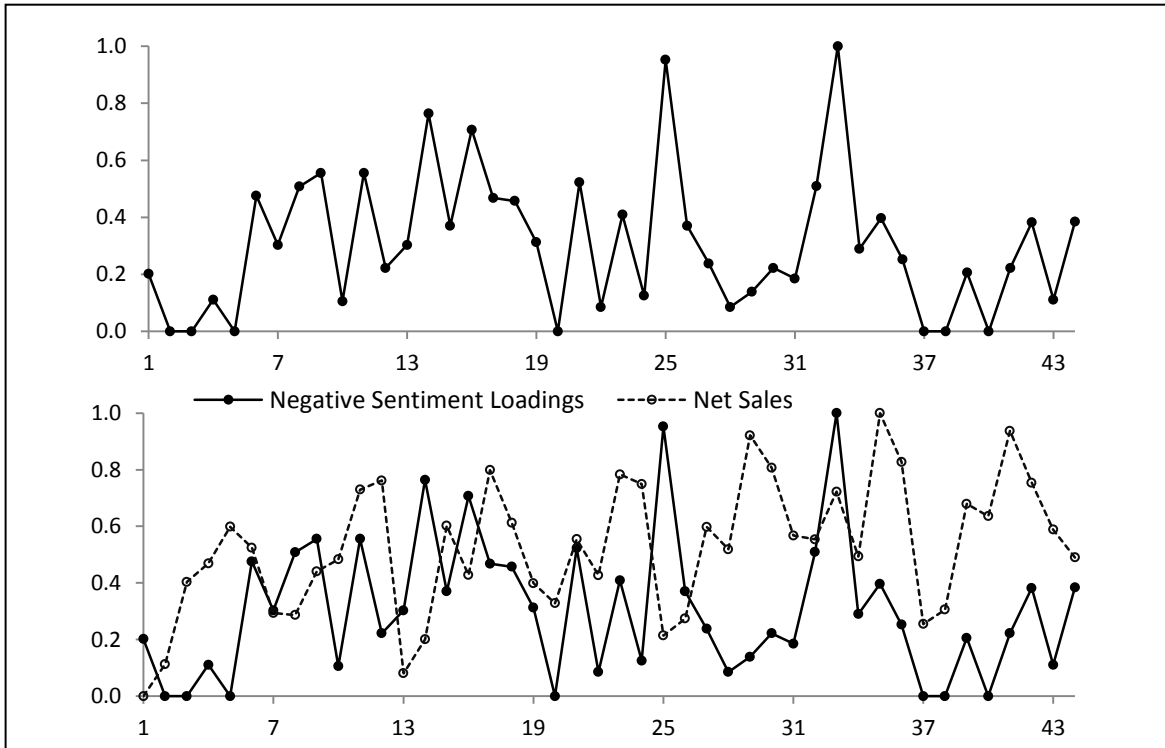


Figure 3. Monthly Time Series of Negative Sentiment Data

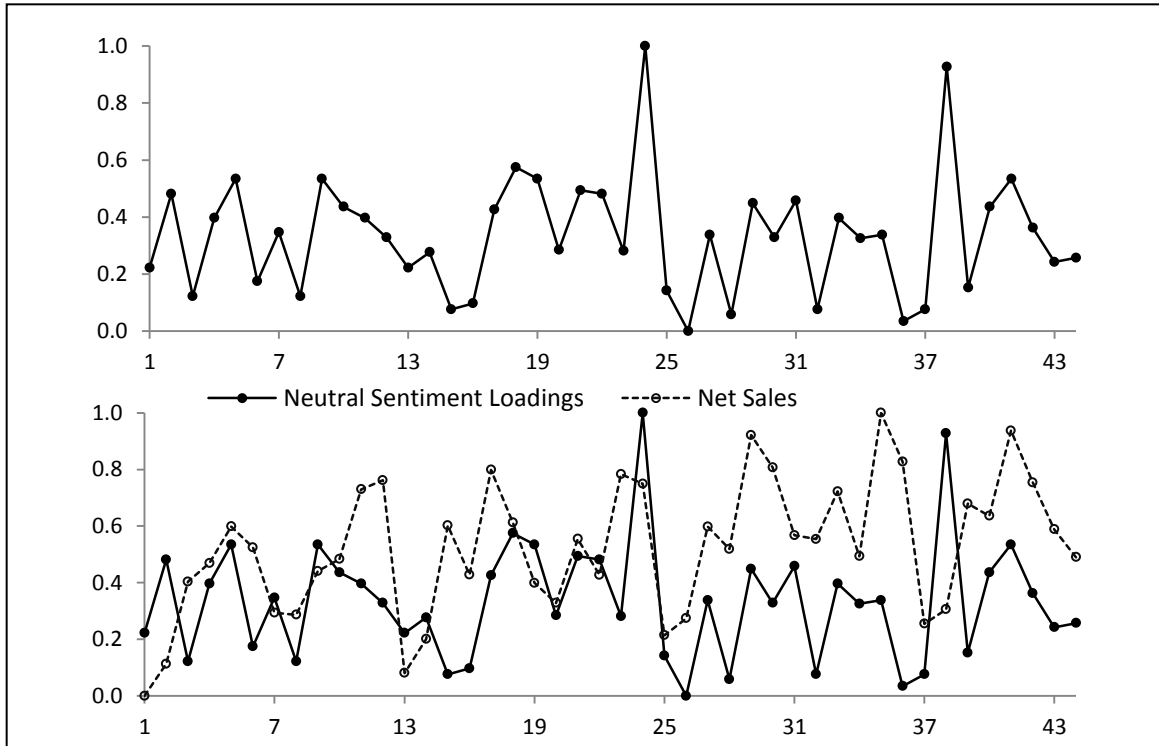


Figure 4. Monthly Time Series of Neutral Sentiment Data

An initial summary of the data in this analysis is shown in Table 2. The summary statistics for the sentiment data are based on the N = 887 tweets about the retail brand collected from twitter online. Here, for example, the mean positive sentiment is about 0.2 whereas the same statistics for the negative and neutral sentiments are 0.27 and 0.37, respectively. Table 2 also shows about 57% of these tweets were classified as positive, 37% were classified as negative. About 14% of all tweets were evaluated to have neutral sentiments. In general, these results indicate that on average, tweeters have a favorable view of this brand. The standard deviations of these measures are relatively large, ranging between 0.15 and 0.48, suggesting a plurality of sentiments about this brand.

Table 3 shows the correlation of the sentiment data, and the demographic and economic data and their relationships with the retail sales. The positive, negative and neutral sentiment variables have low correlation to the retail sales, with their correlation of -0.28, 0.02 and 0.20, respectively. The positive sentiment is negatively correlated with the net sales. This initial finding is surprising since one would expect positive feeling about a brand be associated with an increase in customer spending behavior. This negative correlation could be an artifact of the concurrency of the positive sentiment and the sales of the brand. We have more to say about this finding in the subsequent sections of this paper.

Also, retail sales tended to correlate more highly with the demographic variables than with either the sentiment or economic variables. Although the sentiment variables seem to relate weakly with retail sales (the outcome variable), these relationships will be investigated further in the following sections.

Statistics	Factor 1 (Positive Sentiment)	Factor 2 (Negative Sentiment)	Factor 3 (Neutral Sentiment)
Mean	0.20	0.27	0.37
Standard Deviation	0.15	0.17	0.48
Non-zero value Tweets	507 (57%)	332 (37%)	127 (14%)
Zero-value Tweets	380 (43%)	555 (63%)	760 (86%)
Total Tweets	887 (100%)	887 (100%)	887 (100%)

Table 2. Statistics of Sentiment Variables

Statistics	Retail Sales	Recency	Age	Tenure	Yield Rate	Positive Sentiment	Negative Sentiment	Neutral Sentiment	
Mean	0.53	0.50	0.53	0.59	0.41	0.36	0.31	0.34	
Standard Deviation	0.24	0.25	0.31	0.35	0.27	0.20	0.25	0.21	
Correlation Matrix	Retail Sales	1.00	0.39	0.46	0.46	-0.16	-0.28	0.02	0.20
	Recency		1.00	-0.29	-0.21	0.10	-0.14	-0.24	0.17
	Age			1.00	0.88	-0.07	0.03	-0.03	0.00
	Tenure				1.00	-0.03	0.02	-0.12	0.04
	Yield Rate					1.00	0.29	-0.28	-0.14
	Positive Sentiment						1.00	-0.06	-0.59
	Negative Sentiment							1.00	-0.22
	Neutral Sentiment								1.00

Table 3. Example of Sentiment Loadings

5. MODEL HYPOTHESES

A number of trends affect retail business success. Some of these trends are more conventional and tie more closely to economic growth, credit availability and demographics. Other trends affecting retail business are relatively new, poised to become a mainstay, and are closely tied to technology (e.g., mobile marketing and social media). An understanding of these new trends might afford companies greater insight of the impact of that, for example, social media may have on customers/brand relationship.

At the simplest level, we posit a relationship exists between retail performance (a single measured outcome variable) and sentimentality and other measured variables. This simple hypothesis can be tested within a multiple regression framework. Since we are dealing with time series data, however, we opt for an autoregressive model to correct for serial correlation. The AUTOREG procedure in SAS augments the linear regression model with an autoregressive model to account for the autocorrelation of errors. It simultaneously estimates the regression coefficients and the autoregressive error model parameters. The AUTOREG procedure in SAS can fit autoregressive models of any order (SAS Institute Inc., 2014).

In this paper, we present a series of models using the PROC AUTOREG procedure. We start with a pure autoregressive model which is built for the retail sales of a brand without adding any extra predictive variables. Then we add customer sentiment variables to see if the sentiment variables can be used to predict the business performance. Then we add the extra predictive variables including demographic and economic variables to the model. This incremental modeling process is demonstrated in the following results section.

6. RESULTS

6.1 SENTIMENT ANALYSIS ACCURACY

As expected, the survey results showed that human raters tended to agree about 70% percent of the time. The fact that agreement is not 100% is a strong indication that there is significant variation in how people interpreted and evaluated these tweets. A comparison of the human ratings and the sentiment classification system showed that the scores were very similar. The classified sentiments for each tweet were compared to the human ratings by checking if the classified sentiment score is within the one standard deviation range of the mean sentiment of human rated scores.

We obtained the following results: 1) there was about a 66% agreement between the humans and our classification system for the positive sentiment dimension; 2) there about a 74% agreement between the humans and our classification system for the negative sentiment dimension; 3) there about a 70% agreement between the humans and our classification system for the neutral sentiment dimension.

Based on these results, we concluded that the sentiment classification system did a relatively good job at replicating human judgment when it comes to classifying twitter sentiments to this retail brand.

6.2 SENTIMENTALITY LAGS TESTING

When we test the relationship of the customer sentiment to the retail sales, we need to consider how far in terms of time the sentiment variable will be related to the retail sales. We first test the length of the separation time using the SAS REG procedure. The test was performed on each of the three sentiment dimensions - positive, negative, and neutral sentiments - separately. A separation time up to 6 months from the month of interest is considered in the test. Stepwise regression was performed on the retail sales using sentiment variables with time lags of 0, 1, 2, 3, 4, and 5 months. Significant level at 0.1500 is used to select variables during stepwise regression. The results showed that all the negative and neutral sentiment variables with time lags of 0, 1, 2, 3, 4, and 5 months were not significant. The positive sentiment with time lags of 1, 3, and 4 months were identified as significant variables (see Table 4).

Variable	Estimate	Standard Error	Type II SS	F value	p-value
Intercept	0.46058	0.10033	0.86129	21.07	<.0001
Positiv_lag0	-0.4357	0.16383	0.28906	7.07	0.0117
Positiv_lag3	0.39391	0.16216	0.24117	5.9	0.0204
Positiv_lag4	0.27992	0.16198	0.12206	2.99	0.0928

Table 4. Results of Sentimentality Lags Testing

6.3 SEPARATE AUTOREGRESSIVE TERMS MODELS

The first model fitted is an autoregressive model with autoregressive (auto-lag) terms only. This model assumes that variation in retail sales performance over time is solely a function of prior sales behavior, plus random errors at each time of evaluation. Autocorrelation test is first performed on the retail sales data. Strong autocorrelation was identified at time lags of 1, 6, 12, and 13 months. Then autoregressive modeling was performed on the retail sales using the identified time lags. The modeling results showed that the significant autoregressive terms were the retail sales with the time lags of 1, 12 and 13 months. The initial autoregressive model M0 was generated by placing these optimal parameters in the autoregressive model. The model leads to a high Total R-Square = 0.8693, DFE (Degree Freedom of Error) = 36, RMSE (Root of Mean Squared Error) = 0.08333. The detailed model results are shown in Table 5. The results indicate that this highly restrictive model does fit these data very well.

6.4 ADDING SENTIMENTALITY INFORMATION

The second set of models evaluated is an autoregressive model which, in addition to the initial autoregressive terms and unique error, examines the significance of the three sentiment dimensions – positive, negative and neutral sentiments - individually and collectively. The above sentimentality lags testing results showed that only the positive sentiment variables with time lags of 1, 3, and 4 months are significant. Autoregressive modeling by adding these identified sentiment variables with time lags led to the model M1 in Table 5. The model includes the significant variables with a p-value less than 0.10. We can see that one sentiment variable, the positive sentiment with a lag of 3 months, was kept in the model. The significant level for this variable is 0.05. The Total R-square = 0.8866, and RMSE = 0.07874. The Regression R-square without the autoregressive terms is 0.1067. It indicates that the sentiment variable in the model, the positive sentiment with the time lag of 3 months, is a significant variable to affect the retail sales.

6.5 ADDING DEMOGRAPHIC AND ECONOMIC INFORMATION

In addition to the initial autoregressive terms and the sentiment variables, we added demographic and economic variables into the autoregressive model. These variables include customer tenure, customer age, customer recency, and the treasury 10-year yield rate as we described in the data description section. The models M2, M3, and M4 in Table 5 are the modeling results by adding these demographic and economic variables. Customer tenure and customer age were found to be significant when we added each of them to the model M1 separately. The details of these two models are shown in the models M2 and M3 in Table 5. However, customer recency and 10-year treasury yield rate were not significant when we added them separately into the model M1. The combination of variables was tested too by adding any 2, 3 or 4 of these four variables into the model M1. The model M4 in Table 5 is the model by adding customer recency and customer age in addition to the initial autoregressive terms and the identified sentiment variable. All of the four estimated coefficients are significant in the model M4, and noteworthy. Although these independent variables are correlated, the combination of both of the age and recency produces additional changes in the outcome. The amount of additional variance accounted for due to the demographic and economic variables is 4.65% (Total R-Square = 0.9331). The sentiment variables remain fairly stable in all of these four models from M1 to M4 in Table 5.

Model	Variable	Parameter Estimates				Model Fit Index			
		Estimate	Standard Error	t value	p-value	DFE	RMSE	Regression R-Square	Total R-Square
M0	Intercept	0.4752	0.0851	5.59	<.0001	36	0.08333	0.0000	0.8693
	AR1	-0.0273	0.0508	-0.54	<.0001				
	AR12	0.1231	0.0583	2.11	<.0001				
	AR13	0.0400	0.0546	0.73	<.0001				
M1	Intercept	0.4919	0.0828	5.94	<.0001	35	0.07874	0.1067	0.8866
	Positiv_lag3	0.0958	0.0476	2.01	0.0519				
	AR1	-0.7024	0.1222	-5.75	<.0001				
	AR12	-0.9125	0.0520	-17.54	<.0001				
	AR13	0.7363	0.1363	5.40	<.0001				
M2	Intercept	0.4112	0.0836	4.92	<.0001	34	0.07916	0.1797	0.8886
	Positiv_lag3	0.1054	0.0511	2.06	0.0468				
	Tenure	0.1477	0.0778	1.90	0.0661				
	AR1	-0.5671	0.1508	-3.76	0.0006				
	AR12	-0.9110	0.0544	-16.75	<.0001				
	AR13	0.6012	0.1653	3.64	0.0009				
M3	Intercept	0.3518	0.0686	5.13	<.0001	37	0.07317	0.525	0.8969
	Positiv_lag3	0.0928	0.0535	1.73	0.0914				
	Age	0.2781	0.0459	6.06	<.0001				
	AR12	-0.9405	0.0399	-23.6	<.0001				
M4	Intercept	0.4757	0.0905	5.26	<.0001	35	0.06048	0.6693	0.9331
	Positiv_lag3	0.1111	0.0446	2.49	0.0177				
	Age	0.2367	0.0432	5.48	<.0001				
	Recency	-0.2151	0.0773	-2.78	0.0087				
	AR12	-0.9702	0.0220	-44.04	<.0001				

Table 5. Autoregressive Models with Extra Predictors

7. DISCUSSION

7.1 SUMMARY

We presented a systematic classification approach to convert unstructured social media data into structured customer sentiment data. We then presented several statistical models to explain the relationship between sentimentality and retails sales performance while controlling for some other key drivers of sales. We used data from a sample of more than thousands of customers who made at least one purchase during 2011 January to 2014 August at a retail brand. We used twitter posts about a particular brand from 2011 January to 2014 August. We also used demographic and economic data collected in that same period of time. Autoregressive models were fit with sentiment scores, demographic, and economic variables.

Our main findings can be outlined as follows:

1. The classification system we built to convert the unstructured twitter text data to customer sentiments was very accurate. The system matches human judgment at a rate of 70% across all three sentiment dimensions: positive, negative and neutral sentiments.
2. Only the positive sentiment was found to be a meaningful predictor of retail sales, and remains significant even after accounting for autoregressive error terms, and other sales driving covariates including demographic and economic variables.
3. The initial correlation between the positive sentiment variable and retail sales was negative. However, this correlation turned to be positive when time lags and additional covariates were added to the model.

7.2 CONCLUSION

This paper presented a novel way to classify social media data to be used for predicting retail sales performance. We used the unstructured media data about a retail brand posted on twitter along with the traditional data including the demographic, economic and transaction data. A systematic classification method was built using SAS to convert the unstructured twitter text data into structured customer sentiment data. By classifying each tweet message to have three sentiment scores in positive, negative and neutral sentiment dimensions, this method keeps more sentiment information than using only one sentiment score for each tweet. The classification results showed that this modest classification system worked well to match the human ratings at a rate of 70%.

This paper built sets of models using SAS procedure PROC AUTOREG to test if sentiment variables are significant to predict retail sales performance. The results have some potentially important implication. The autoregressive models show that sentiment scores are moderately related to the retail sales performance. Even after we added other traditional variables to predict retail sales, sentiment variable still remained significant with meaningful estimates. The overall model predications we produced are moderate but these variables do help cut down on the uncertainty about the utility of customer sentiment.

8. REFERENCES

- Davenport, T. H. and Dyché, J. 2013. "Big Data in Big Companies". Thomas H. Davenport and SAS Institute Inc.
- Helweh, A. 2011. "The Future of Sentiment Analysis". Retrieved August 30, 2004, from <http://www.socialmediaexplorer.com/social-media-monitoring/sentiment-analysis/>
- Loehlin, J. C. 1992. "Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis" (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum.
- Mercier K., Richards B. and Shockley R. 2013. "Analytics: The real-world Use of Big Data in Retail". IBM Corporation.
- Ogneva, M. 2010. "How Companies Can Use Sentiment Analysis to Improve Their Business". Retrieved August 30, 2004, from <http://mashable.com/2010/04/19/sentiment-analysis/>
- Troester, M. 2012. "Big Data Meets Big Data Analytics". Cary, NC: SAS Institute Inc.
- SAS Institute Inc. 2014. "SAS/ETS® 13.2 User's Guide The AUTOREG Procedure". Cary, NC: SAS Institute Inc.
- Schroeck, M., Shockley R., Smart, J., Romero-Morales D., and Tufano, P. 2012. "Analytics: The real-world Use of Big Data". IBM Corporation.
- Cambria, E., Schuller, B., Xia, Y., and Havasi, C. 2013. "New Avenues in Opinion Mining and Sentiment Analysis". IEEE Intelligent Systems 28 (2): 15–21.

9. ACKNOWLEDGMENTS

I would like to express my sincerest thanks to Dr. Alphonse Damas for his unlimited encouragement and guidance through this project. I also want to thank Tim Sweeney and Amy Yu for their great support to this project. Special thanks are due to Christian Billman, Uchenna Chilaka, and Ben Elbert for their continued help on data collection. I finally thank the Marketing Insights Department at Alliance Data Systems, Inc. for providing the opportunity to work on this paper.

10. CONTACT INFORMATION

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

Name: Li Zhang
Enterprise: Alliance Data Systems, Inc.
Address: 3100 Easton Square PI
City, State ZIP: Columbus, Ohio, 43219
E-mail: li.zhang@alliancedata.com

11. APPENDIX

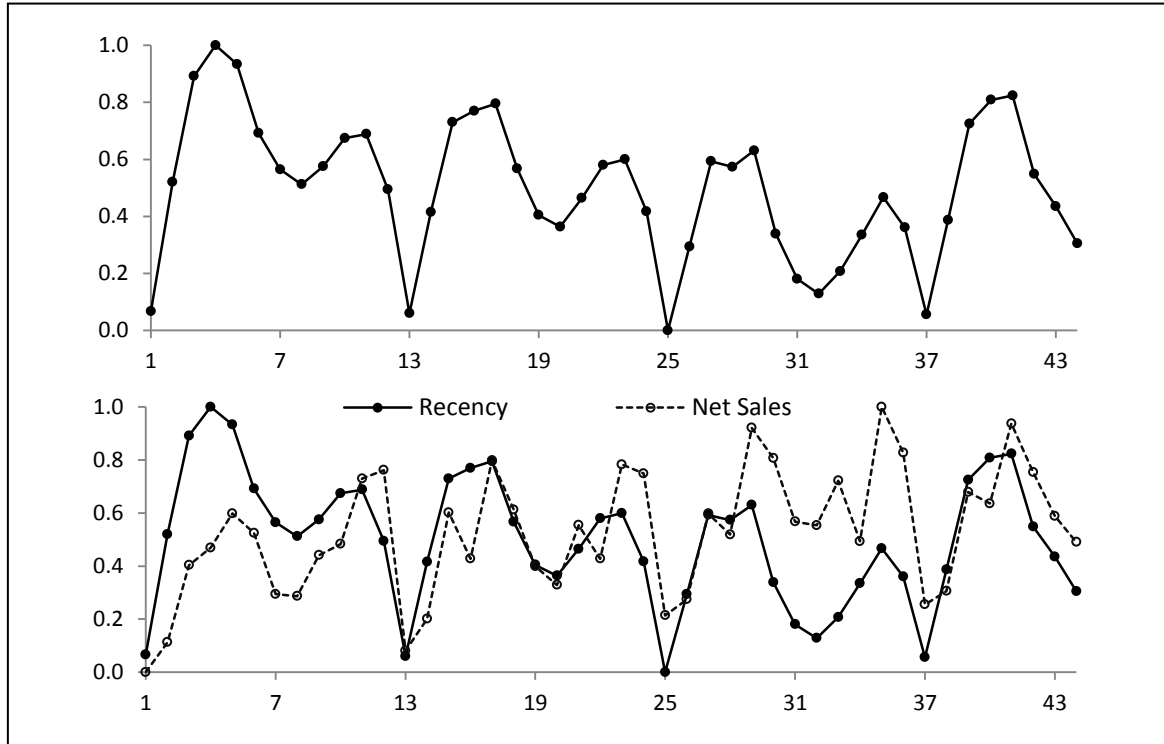


Figure 5. Monthly Time Series of Customer Recency

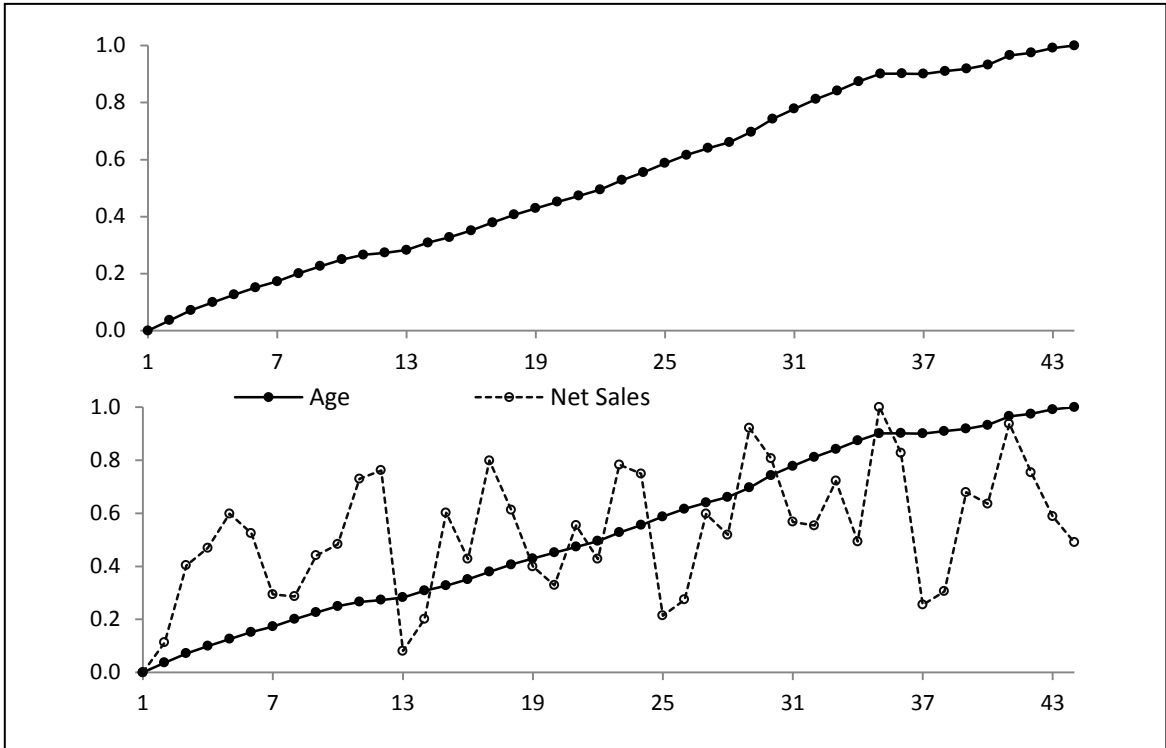


Figure 6. Monthly Time Series of Customer Age

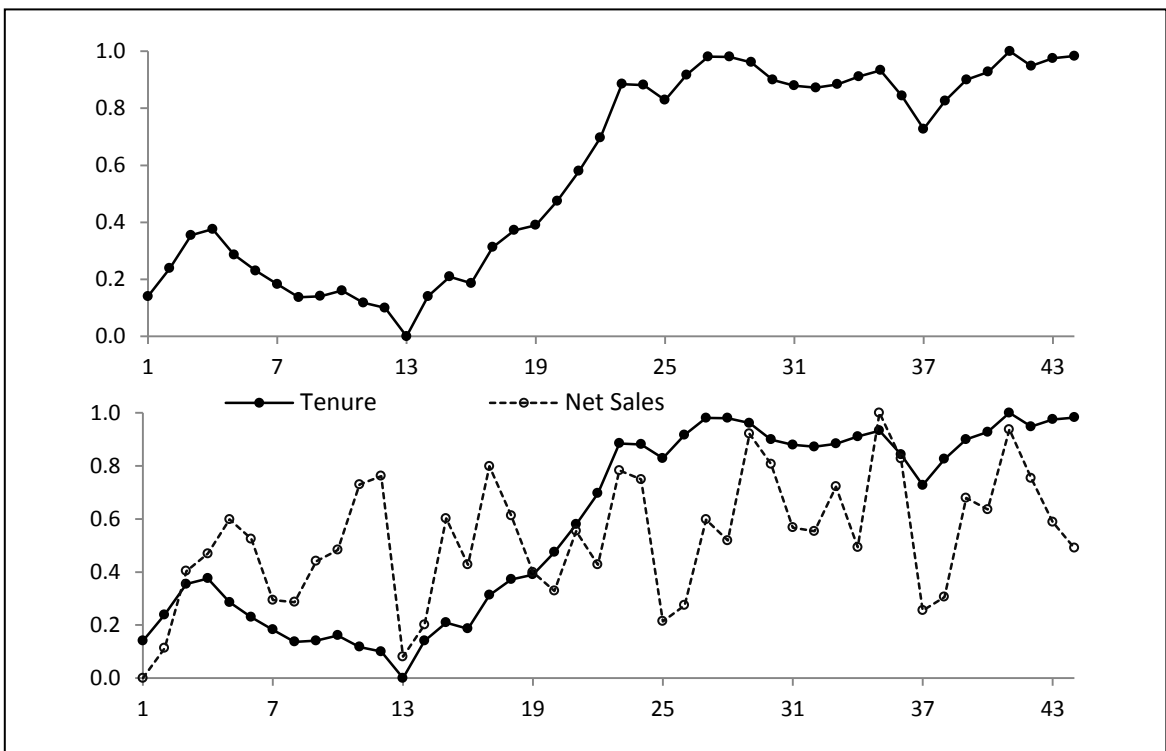


Figure 7. Monthly Time Series of Customer Tenure

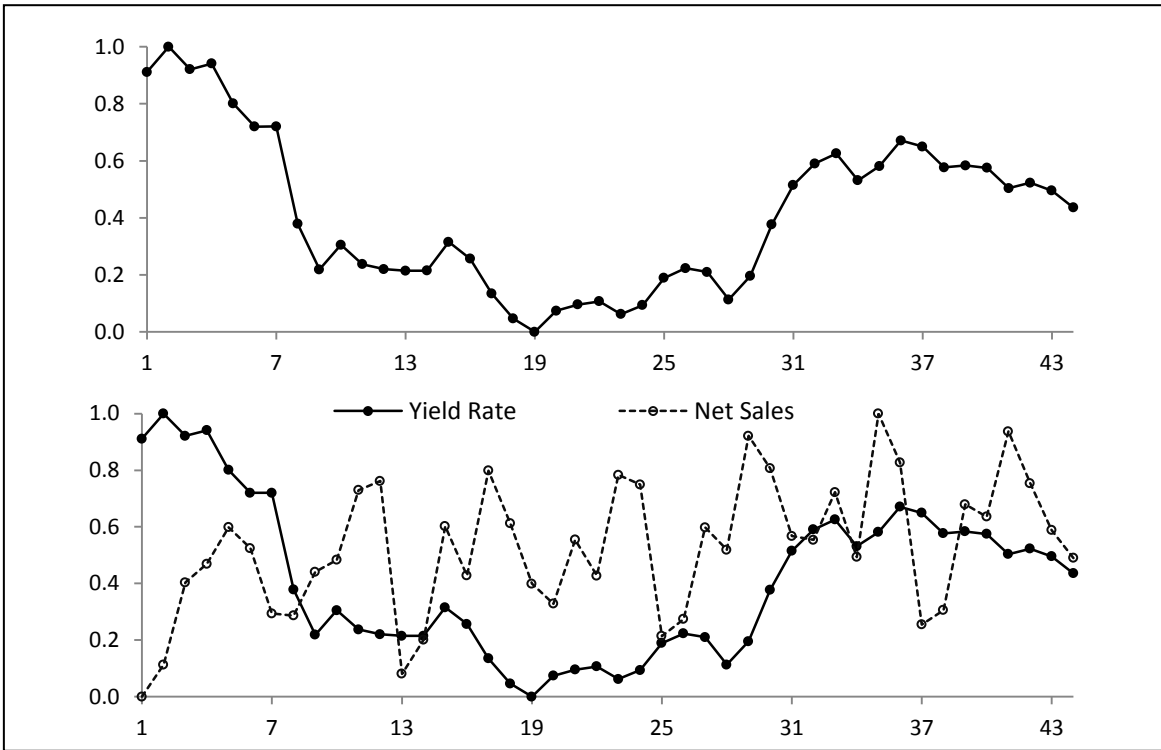


Figure 8. Monthly Time Series of Treasury 10-year Yield Rate

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.