

Predicting Smartphone Sales from Online Product Review Sentiments

BYGB 7978 Web Analytics

Instructor: Professor Yilu Zhou

16 December 2015

Section 1 - Group 1

Nikki Alcantara

Hang Sun

Nan Wang

Jiahui Xu

Executive Summary

A recent study was published in an online article on Social Times (*Morrison, K. 2014*) highlighting that 81% of shoppers conduct online research before buying which is evident in the advent of countless professional product review sites, unboxing videos on YouTube, and customer reviews. By 2019, the number of mobile device users is expected to grow to as much as 14.8 billion worldwide, with each business user owning an average of 2.2 devices (*The Radicati Group, Inc., 2015*). That brings in a total of 12 billion people researching on mobile devices, reading the opinions of other buyers and users online. It is, therefore, worth asking whether the sentiments expressed online can actually predict sales volume of these mobile devices.

The group focused on smartphones, specifically on Samsung Galaxy Phones and Apple iPhones. Methods of analyses include sentiment analysis, time series trend analysis, and comparative linear regression between a model without sentiment and with a sentiment factor to predict sales volume. The results show that in both brands, an improvement in the regression model is observed in functions wherein the count of positive comments is included. This shows that smartphone manufacturers must pay close attention to the comments posted by the general public related to their products as these affect sales to some extent. By monitoring online sentiments, impacts of negative sentiments can be mitigated while positive comments can be leveraged to increase sales.

It must be noted, however, that the financial data for Samsung was not perfectly consistent with the other variables as it consisted of the sales of their other mobile devices apart from Galaxy smartphones. Since the scope covers only a short period of time and comments from only one website, further research must be also be conducted.

Business Goal Analysis

The availability of advanced data mining techniques alongside traditional statistical methods such as regression and trend analysis have enabled people to perform accurate sales predictions in a more structured manner. Sentiment analysis, however, has not been extensively used as a component in sales predictions. The goal of this study, therefore, is to demonstrate the strength of online sentiment in improving sales volume predictions.

The popularity of social media such as forums and review sites have changed the way people interact, making it easier to voice out product feedback and shopping experiences. Since all these data is available online, web analytics provides the necessary techniques to collect and mine these data to assist in important business functions such as sales predictions. Sales volume can be affected by numerous factors which makes it difficult to find the optimal prediction model. In this project, the group plots the sales trends of Apple iPhones and Samsung Galaxy phones against results of sentiment analysis of comments related to the said products. Despite the simplicity of the method, it provides immediate valuable insight into their correlations and thus gives the companies better understanding of their sales' trends in comparison with their products' ratings in terms of the amount of positive/negative comments online. Additionally, by employing regression analysis, the relationship between online sentiment and sales volume can be numerically proven.

The results of this study will assist businesses in optimizing sales and distribution, inventory, manufacturing, and most importantly, online marketing strategies. With more advanced text analytics and data mining, businesses will be able to further assess the product's popularity and reputation as well as to help determine future product improvements.

Dataset Description

From online product review website GSMArena, the group used a Python program to obtain, for each review, the post type, username as well as the date posted and content. After missing values have been cleaned, the dataset overall contains 35,963 user reviews for iPhone models and 51,539 user reviews for Galaxy models. GSMArena is well-known for the popularity of smartphone reviewing around the globe. The users, therefore, are from a variety of backgrounds which leaves the pool less biased. Meaning Cloud, an efficient sentiment tool, was used to give sentiment classification for each review content. The resulting categories are P+, P, N+, N, neutral and null. To simplify, the group combined P+ and P as positive sentiment and N+ and N as negative sentiment. The group then summed the count of positive reviews and negative reviews on quarterly basis for the trend and regression analyses.

For the financial data, Apple and Samsung's publicly available quarterly financial reports from 2011 to 2015 were used. The dataset includes smartphone units sold, revenue from smartphone sales for Apple, revenue for Samsung's IT and Mobile Communications segment (IM), Apple's Research & Development expenses, and Samsung's advertising expenses.

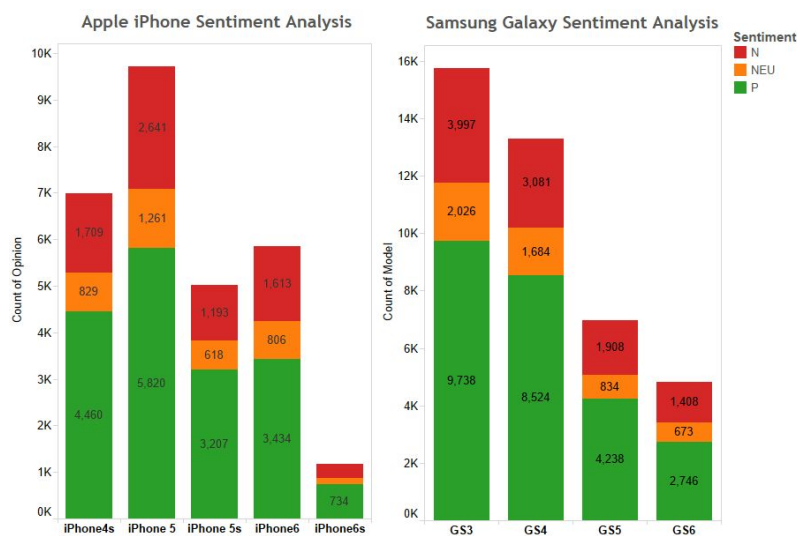
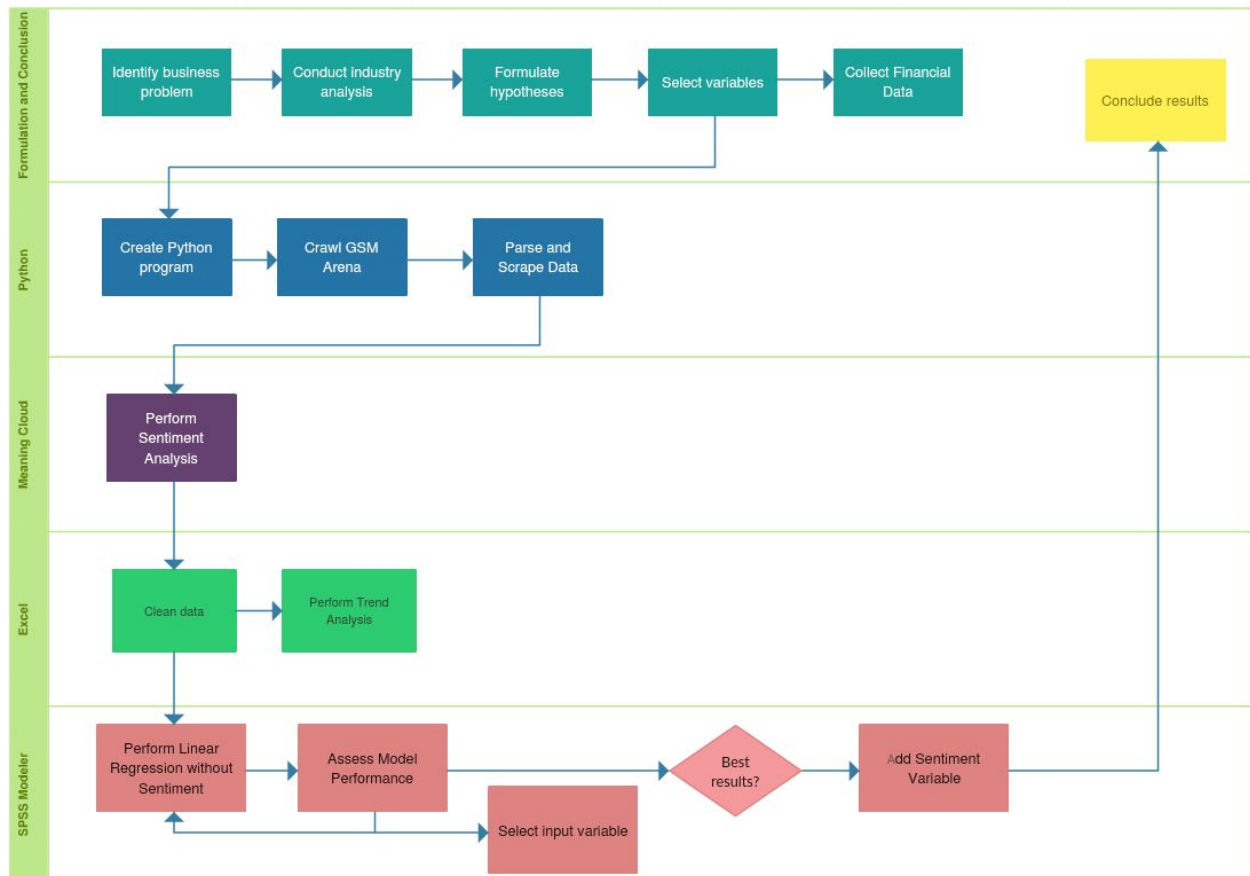


Figure 1 - Sentiment Analysis Results by iPhone Model

System Design



First of all, the group identified the business problem as that of a growing number of consumers researching online prior to purchasing any product, so the association between sales and review sentiment was worth exploring. By conducting a high-level industry analysis, it was decided that the best candidates are Apple and Samsung, taking the leading positions in the market with the highest sales of smartphones worldwide. The group then decided to limit the scope to iPhones 4s, 5, 5s, 6, 6s and Samsung Galaxy S3, S4, S5, S6 models due to their popularity. From this, the group hypothesized that the number of positive sentiments of online product reviews are highly positively correlated with sales volume. With some research, it was determined that GSMArena was the ideal website with thousands of smartphone reviews available for extracting using

python program importing the Beautiful Soup API. From these nine smartphone models alone, 5,176 pages were crawled to collect 35,963 user reviews for iPhone and 51,535 for Samsung Galaxy phones. The same program automatically parsed and scraped the output to structure the required data of four variables: post type (reply to a previous comment or a new post), date posted, username and opinion into csv files. Since the output was stored in different files, the group consolidated these into two: one for iPhone and another for Samsung and proceeded with cleaning the data. Meaning Cloud API in Excel was then used to perform the sentiment analysis on all these user reviews. This software classified each user review/comment according to six categories: P+, P, N+, N, neutral and null. To simplify the result, the group regarded both P+ and P as a single value of P and did the same for negative sentiments: N+ as N. The group therefore only has four classified groups concerning sentiment, which are positive, negative, neutral and null. Additionally, the data on sales revenue, advertising expense, R&D expense and units sold on a quarterly basis from 2011 Q4 - 2015 Q4 were obtained from Apple and Samsung financial statements. Now with a complete dataset, descriptive trend analysis between sales volume and the number of reviews per sentiment category was done in Excel. Despite the visibility of clear trends for Apple and the lack of observation for Samsung, the group used SPSS Modeler to create two regression models to be able to numerically measure and compare the impact of the count of positive sentiments as a factor in predicting sales volume. After several iterations of different models, the group found that a good comparison can be made when the first model has the previous year's sales volume as input and the second model has the same input with the addition of the number of positive comments during the previous quarter. To further illustrate, the sales volume of Q3 of 2014 would be plotted against Q3 2013 sales volume as the

non-sentiment input and Q2 2014 number of positive reviews as the sentiment factor in the model. It must also be understood, that this was not done per phone model, but was aggregated on a company-level. Sales volume for one quarter, for instance, encompasses all sales from all iPhones available during the period. The reviews are also aggregated regardless of the phone model and is grouped according to the date posted on the website. The results of two models have been carefully compared and it was apparent that in both Apple and Samsung studies, the addition of the count of positive reviews resulted to an improved regression model as measured by the R and R-squared figures. Details of the implementation and evaluation of the results will be discussed further in the succeeding sections of this paper.

System Implementation

Trend Analysis

In order to see the trends and correlations directly, the group plotted the quarterly sales volume, number of reviews (quarterly) and number of positive reviews (quarterly) for both iPhone and Galaxy. For both smartphones, the number of reviews and number of positive reviews have very high positive correlation because the number of positive reviews is always approximately 50% of the total.

An interesting pattern of iPhone was discovered in that the sales volume always followed the number of positive reviews in the previous quarter. When the number of positive reviews in Q1 increased, the sales volume would increase in Q2; when the number of positive reviews in Q2 decreased, the sales volume would consequently drop in Q3. Thus, the number of positive reviews is found to be a good predictor for the sales volume in the following quarter. In order to

show the correlation between current quarter sales volume and previous quarter sentiments, a graph depicting the trends across a four year span (2012 - 2015) with current quarter sales volume, previous quarter number of reviews and number of positive reviews was modeled. The graph for iPhone showed an obvious pattern that these three variables always increased and decreased together. The same pattern, however, was not observed with the Samsung Galaxy dataset.

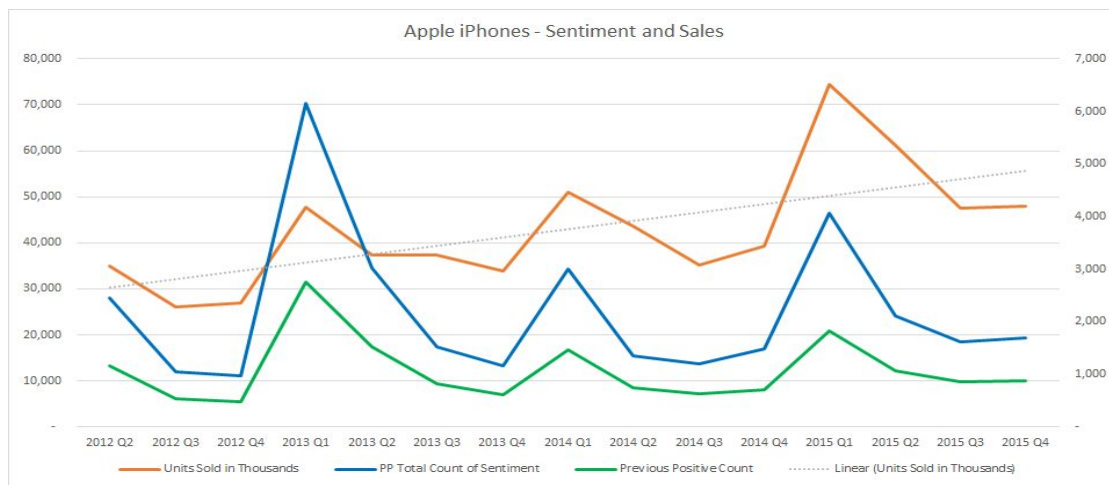


Figure 2 - Apple iPhone Trend Analysis - Sentiment and Sales

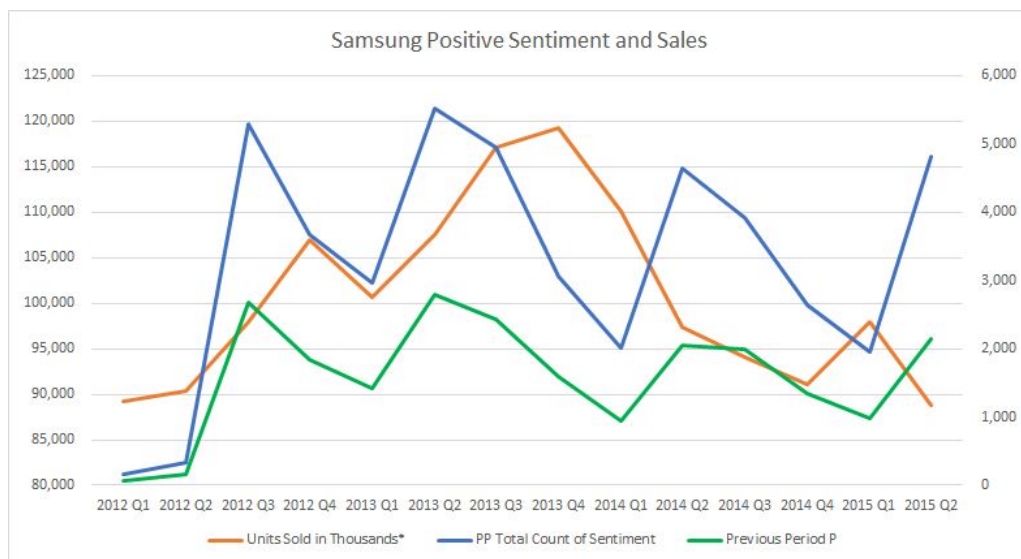


Figure 3 - Samsung Galaxy Trend Analysis - Sentiment and Sales

Regression Modeling

Since the review sentiments is a possible predictor of sales volume as seen in the trend analysis, the group wanted to further explore statistically whether it did improve sales performance. A multiple regression model was designed to predict sales volume to assess whether the model would perform better by adding sentiment as an independent variable in addition to the previous year, same quarter sales. The model for iPhone without any sentiment variable showed a clear cycle in the sales volume of iPhone in which whenever a new model was released (every September, at the end of Q4), the sales would significantly increase, followed by three succeeding quarters of gradual drop in sales. Due to this annual cycle, quarterly sales volume in previous year was selected as the predictor (independent variable) of the sales volume in the same quarter in the following year (dependent variable). For example, sales volume in 2012 Q2 was used to predict 2013 Q2. The model was simple but had a good performance: R Square=0.744, Adjusted R Square=0.725.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.863 ^a	.744	.725	.524684

a. Predictors: (Constant), Previous Year's Units Sold_transformed

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.540E-15	.135		.000	1.000
	Previous Year's Units Sold_transformed	.863	.140	.863	6.153	.000

Figure 4 - Apple iPhone Sales Volume - Regression Model without Sentiment

A sentiment variable was included in the previous model to test whether adding this variable into the equation would result to a better performing regression model. After running the model with

the two independent variables, previous year quarterly sales volume (same as the previous model) and number of positive reviews in previous quarter, this new model showed an increased R Square and Adjusted R Square: R Square=0.766, Adjusted R Square=0.727.

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.875 ^a	.766	.727	.522800	

a. Predictors: (Constant), Previous Positive Count_transformed, Previous Year's Units Sold_transformed

Coefficients					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	3.455E-15	.135		.000
	Previous Year's Units Sold_transformed	.785	.158	.785	.000
	Previous Positive Count_transformed	.166	.158	.166	.316

Figure 5 - Apple iPhone Sales Volume Regression Model with Sentiment

Same process was done for the Samsung Galaxy dataset. Quarterly sales volume in previous year and number of positive reviews in previous quarter were used as inputs for predicting the sales volume in the same quarter in the following year. The model without sentiment variable had R Square=0.035 and Adjusted R Square=-0.045. After adding the sentiment variable, the new model had R Square=0.134 and Adjusted R Square=-0.023.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.187 ^a	.035	-.045	1.022431

a. Predictors: (Constant), Previous Year's Sales_transformed

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.286E-14	.273		.000	1.000
	Previous Year's Sales_transformed	.187	.284	.187	.660	.522

Figure 6 - Samsung Galaxy Sales Volume Regression Model without Sentiment

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.367 ^a	.134	-.023	1.011449

a. Predictors: (Constant), P_transformed, Previous Year's Sales_transformed

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.389E-14	.270		.000	1.000
	Previous Year's Sales_transformed	.088	.294	.088	.301	.769
	P_transformed	.330	.294	.330	1.123	.285

Figure 7 - Samsung Galaxy Sales Volume Regression Model with Sentiment

The performance of both models were not as good as iPhone, but after adding the sentiment variable, the new model still got higher R Square and Adjusted R Square. Thus, this research has sufficient models to conclude that adding the sentiment variable improved the performance of sales volume prediction models for both iPhone and Samsung Galaxy smartphones.

Evaluation

For Apple iPhones and Samsung Galaxy smartphones, the results of the study has shown positive correlation between number of positive sentiment reviews online and the phones' sales volume. With these results, Apple and Samsung should, therefore, constantly align their marketing strategies according to online reviews so that negative comments can be controlled while positive comments can be leveraged to further improve sales. An example could be through online promotional strategies targeted towards cyber citizens to encourage more pleasant reviews on their products. Also, since the amount of positive comments is positively correlated with sales performance before and during the product release date, the group recommends that advertising

and marketing expenses be increased during this period to further boost sales.

Conclusion and Future Direction

The regression model used for sales prediction do perform much better with the addition of sentiment analysis, specifically the amount of positive comments, which translates to improved sales prediction model. To be specific, positive views or comments in the previous period are highly likely to indicate an increase in sales in the following quarter. Some limitations, however, include 1) having only a single website as source for the opinions which may be biased to a specific set of users; 2) brands and models were only Apple and Samsung Galaxy wherein trends may be different for other smartphones; and 3) only one sentiment analysis tool was used and validation/comparison was not conducted for another tool. Recommendation for future research should address these by expanding the dataset to more websites, brands, and models, by extending the periods from earlier than 2011 and later than 2015, and by using several sentiment analysis tools. Lastly, text analytics may be done to explore further patterns in the user reviews.

References

Apple Investor Relations Financial Highlights. Retrieved from:

<http://investor.apple.com/financials.cfm>

Liu, M. Hu, & J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings WWW 2005, pages 342–351, 2005.

Number of mobile phones sold* by Samsung worldwide from 2010 to 2015 (in millions), by quarter. Retrieved from: <http://www.gartner.com/newsroom/id/2996817>

Morrison, K. (2014, November 28). *81% of Shoppers Conduct Online Research Before Buying*.

Retrieved from

<http://www.adweek.com/socialtimes/81-shoppers-conduct-online-research-making-purchase-info-graphic/208527>

Pang, L. Lee, & S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings EMNLP 2002, 2002.

The Radicati Group, Inc. (2015, February). *Mobile Statistics Report, 2015-2019*. Retrieved from

<http://www.radicati.com/wp/wp-content/uploads/2015/02/Mobile-Statistics-Report-2015-2019-Executive-Summary.pdf>

Samsung Financial Highlights. Retrieved from:

http://www.samsung.com/us/aboutsamsung/investor_relations/financial_information/IR_Financial2012_quater.html