Predicting iPhone Sales from iPhoneTweets

Niels Buus Lassen[1], Rene Madsen[1]
[1]Computational Social Science Laboratory
Department of ITM, Copenhagen Business School
Howitzvej 60. 2.14, Frederiksberg, 2000, Denmark
nbl@evalua.dk; Rene.Madsen@infomedia.dk

Ravi Vatrapu[1,2]
[2]Mobile Technology Laboratory
Norwegian School of Information Technology
Schweigaardsgate 14, Oslo, 0185, Norway
vatrapu@cbs.dk

***Abstract*— Recent research in the field of computational social science have shown how data resulting from the widespread adoption and use of social media channels such as twitter can be used to predict outcomes such as movie revenues, election winners, localized moods, and epidemic outbreaks. Underlying assumptions for this research stream on predictive analytics are that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product and that the shared digital artefact that is persistent can create social influence. In this paper, we demonstrate how social media data from twitter can be used to predict the sales of iPhones. Based on a conceptual model of social data consisting of social graph (actors, actions, activities, and artefacts) and social text (topics, keywords, pronouns, and sentiments), we develop and evaluate a linear regression model that transforms iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks. This strong correlation between iPhone tweets and iPhone sales becomes marginally stronger after incorporating sentiments of tweets. We discuss the findings and conclude with implications for predictive analytics with big social data.**

Keywords- data science, computational social science,social data analytics, predictive analytics, iphone sales, iphone tweets, twitter

I. INTRODUCTION

Social media has evolved into vital constituents of many human activities. We share aspects of our lives on Facebook, Twitter, Instagram, Tumblr, and many other social media platforms. The resulting social data is persistent, archived, and can be retrieved and analyzed. Social data analytics is not only informing but also transforming existing practices in politics, marketing, investing, product development, entertainment, and news media.

In this paper, we analyze a complex product that generated a large number of opinions on social media. If social media can be characterized as second life for some, then smartphone has evolved into an extension of human body and mind. The product under analytical consideration, Apple iPhone is one of the best-selling products in history and is associated with large amounts of big data on most social media channels. Our paper demonstrates how Twitter social data can be used to predict the future sales of the Apple iPhone. In particular, we analyze the mathematical relationship between twitter social data and iPhone smartphone sales. Our research question is stated below:

Can big social data predict the sales of smartphones?

Our research hypothesis is that smartphone sales are correlated with tweets and can be predicted on the basis of Twitter data. We adopt the method of Asur & Huberman [1] and examine if the same principles for predicting movie revenue with Twitter data can be used to predict iPhone sales. That is, if a tweet can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. We report and discuss a regression model that can predict iPhone sales with 5-10% average error.

The remainder of the paper is organized as follows. Related work on predictive analytics is reviewed in the next section. Theoretical framework section discusses the AIDA sales funnel model and the Hierarchy of Effects information processing model of advertising. Methodology section discusses twitter data collection and statistical modelling. Results section presents the empirical findings in terms of the regression model. Discussion section offers substantive interpretation of the statistical results and concludes with implications for predictive analytics in particular and computational social science in general.

II. RELATED WORK

We deliberately limit the review of extant literature to empirical work that examined the relationship between social data measures (such as facebook posts/likes/comments/shares, and twitter tweets/re-tweets/mentions/polarity etc.) and real-world business outcomes (revenues, stock price etc.).

A. Social Data & Business Outcomes: Data Science

There has been substantial research work [2-7] in the direction of predicting the stock prices of the companies based on the analysis of content from the online media such as news items, web blogs, twitter feeds. For example, Gavrilov et al., [5] applied data mining techniques on the stock information from various companies by clustering them according to their Standard and Poor (S&P) 500 index, whereas the content from the weblogs is used by Kharratzadeh & Coates [6] to identify the underlying

relationships between the companies to make predictions about the evolution of stock prices.

The most notable papers in this regard is from Asur & Huberman [1] showed that social media feeds can be used as effective indicators of the real-world performance. In their work, they used analysis of hourly rate of tweets about movies, their re-tweets and sentiment polarity to accurately forecast the box-office movies revenue. In fact, their prediction of movie revenues based on the social data measures from twitter outperformed the leading market-based predictions of the Hollywood Stock Exchange. In terms of macro-societal relationships, a research study investigated whether the public mood as measured from large-scale collection of Twitter tweets can be correlated or even predictive of Dow Jones Industrial Average (DJIA) values has been explored by Bollen and Mao [3].

B. Social Media Analytics: Information Systems

Previous literature about social media analytics have focused upon user-generated content (UCG) [8-10], as well as the organizational [11, 12], business intelligence [13, 14], and predictive aspects of social data [15-20]. For example, Zimbra et al., [10] combined sentiment analysis with topic analysis in order to analyze a Wal-Mart discussion forum to improve organizational decision-making. Huber et al., [8] studied how companies can use wall posts and comments on Facebook to stimulate user engagement, while Lin and Goh [9] investigated the co-existence of customers and marketers in order to determine the value of their content on social media. Heath et al., [11] empirically studied how a strategic organizational engagement in social media can advance organizational goals, while Larson and Watson [12] introduced a social media ecosystem to explain the different stakeholder positions in and around the company. Dinter and Lorenz [13] articulated a research agenda for social business intelligence (social BI), while Rosemann et al., [14] sought to advance the conceptual design of BI with data identified from social networks amongst others through a discussion of social customer relationship management (social CRM) and social BI.

There is an elaborate body of work done on predictive analytics. Seebach et al., [18] suggested that companies include data on customer's online search into their IT systems in order to increase their sensing abilities and create a more agile business. vd Reijden & Koppius [21] studied how online buzz predicts actual sales across different phases of a product lifecycle. Geve and colleagues [15] used Google's index of internet discussion forums and Google's search trends to predict sales, while Wu and Brynjolfsson [19] used internet searches to predict housing prices. Zhang and Lau [20] developed a business network-based model to analyze and predict business performances (using the proxies of stock prizes). Nann, Krauss, and Schoder [22] analysed multiple online public data platforms such as Twitter and Yahoo! Finance in order to predict the stock market, while Oh and Sheng [17] analysed the predictive power of micro blog sentiments on stock price directional movements.

In general, we find that most of prior related work in the field employs analytical methods for sentiment analysis of the content (social text analytics) or the social network analysis techniques to study social relationships (social graph analytics). When compared to the prior related work, our approach in this paper is novel in the sense that we use both social graph analysis combined with social text analysis (e.g. sentiment analysis) to compute relationship between the social data (e.g. twitter data) and financial performance (e.g. quarterly revenues) of the companies.

Furthermore, as far as we know, we are the first to use twitter data in measuring the relationship between twitter data and quarterly sales of iPhones. That said, we contribute to the knowledge base by empirically investigating a new domain (smartphone sales), theoretically grounding our analysis in relevant domain theories (AIDA & Hierarchy of Effects, discussed next), and extending Asur and Huberman's [1] model to include seasonal weighting.

III. THEORETICAL FRAMEWORK

In this paper, we build on and substantially extend the method of Asur & Huberman [1] for predicting movie revenue with Twitter data to predict iPhone sales. That is, if a tweet can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. In the next section, we discuss the AIDA and Hierarchy of Effects models in order to delineate the conceptual relationship between users' propensity to tweet and the probability to purchase a product.

A. AIDA

AIDA model stands for *Awareness*, *Interest*, *Desire*, and *Action* and refers to the various stages in a sales process. AIDA was first formulated by Elmo St. Lewis and its original criteria have been subsequently modified to fit technological developments as well as changes in consumer behavior [23]. In terms of the relationship between social data about and sale of an iPhone, the AIDA sales funnel is outlined below.

The first step, *awareness*/attention can result from

- news reading
- friends, colleagues, classmates having the iPhone
- tweets, facebook news, other social media info
- commercials
- seeing the iPhone in use on the metro/bus/train etc

The second step, *interest*/knowledge/liking can result from

- role models having the iPhone
- trying a friend's iPhone
- comparing the iPhone with models from Samsung, Nokia etc. in a mobile phone shop
- reading reviews of phones online including social media

The third step, *desire*/preference can involve

- evaluating iOS vs. Android vs. windows mobile, and forming preferences for what is perceived to be most

easy, intuitive, cool, nerdy, configurable, less app costs, most apps etc.

- social influence processes of identification, conformity etc. [24, 25]
- price/needs/features – nice-to-have vs. need-to-have considerations

The fourth and final step, *action*/conviction/purchase, can lead to

- purchase of the new iPhone or one of its competitors.
- holding on to the old mobile/smartphone for a further period
- opting out of the product category of smartphones all together
- product mention/recommendation/review in face-to-face settings (traditional Word of Mouth) and/or online including social media platforms such as twitter

B. Hierarchy of Effects (HoE)

Hierarchy of Effects (HoE) refers to a family of psychological models that seek to explain human information processing of advertisements [26]. It was first formulated by Lavidge and Steiner [27] and has been the subject of much debate in advertising research [28]. HoE posits a psychological cascade of *cognition*, *affect*, and *behavior* in terms of how advertisements work. According to HoE models, advertisements are processed during the *cognition* phase, leading to the formation of a positive, negative or neutral *affection* which in turn leads to subsequent *behavior*. There are three different orderings of the hierarchy [29]:

- *Learning Hierarchy (C-A-B)* is the typical consumer behavior scenario of learning about a product, forming an opinion, and deciding to purchase it or not.
- *Dissonance Hierarchy (B-A-C)* also known as "buyer's remorse" results when consumers purchase the product first without much deliberation and then have negative experiences of it leading to product awareness.
- *Low-Involvement Hierarchy (B-C-A)* occurs in cases of habitual repurchases owing to brand loyalty (Apple iPhone in our case) and/or product type (for example, bottled water)

Tweets about iPhones can play a role on all three different orderings of the HoE listed above in terms of learning about the product, evaluating one's own experience of it with those of others, and engaging with the product as a brand loyalist by following iPhone related twitter streams. Figure 1, taken from [30], shows the close relationship between the AIDA and HoE models.

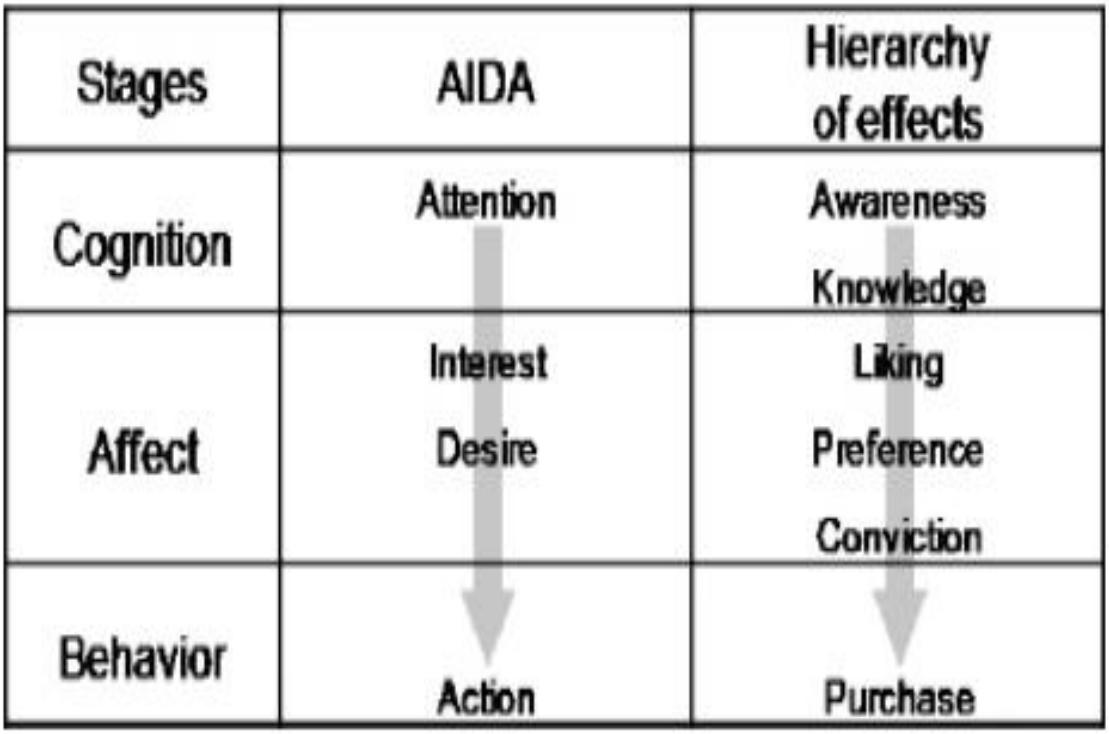


Figure 1: AIDA and Hierarchy of Effects Models

To sum up, tweets about iPhones in particular and smartphones in general are associated with all four stages of the AIDA model and all six stages of the Hierarchy of Effects model. Drawing on Asur and Huberman [1], we treat social data from twitter as a proxy for a user's attention towards the object of analysis which in our case is the iPhone. That said, from the specific domain, we consider a tweet about an iPhone as a proxy for a user's involvement in one of the different stages in the AIDA and HoE models. To be clear, we do not classify each tweet as belonging to a particular stage of AIDA or HoE but treat them as social media manifestations of real-world activities of users/consumers with respect to the iPhone.

IV. METHODOLOGY

A. Dataset

We collected over 400 million tweets containing the phrase "iPhone" in the period 2007-2013 using Topsy Pro Analytics[1]. Technically, our data collection did not use the Twitter firehose, but a Twitter API solution with full access to all Twitter data. We searched for the phrase "iPhone" in Topsy Pro, which then returned number of all tweets (Tweets, retweets, and replies) for the time period specified, and with sentiment numbers calculated. These numbers form the basis for prediction of one quarter sales of iPhones.

We read the numbers of Tweets, and corresponding sentiment number in Topsy Pro on the screen, and inputted those numbers into Microsoft Excel. We employed calendar based quarters rather than the financial quarters of Apple for the modeling.

B. Quantity of Tweets

To provide an example, for the time period of 10-September-2013 to 10-December-2013, we made a data query in Topsy pro, specifying the period and searching for the phrase "iPhone" in all tweets (tweets, replies, retweets). For this example result was 44.62 million tweets and the corresponding sentiment number of 64.

C. Quality of Tweets

The sentiment number in above example expresses 64% of all tweets as positive. The Topsy Pro has calculated this sentiment number on a smaller fraction of the 44,62 mio tweets. The Topsy Pro sentiment algorithm is a black box, and all we know, from their self-reported descriptions, is that it is optimized for English text.

If we define

- *p : Tweets with positive sentiment*
- *n : Tweets with negative sentiment*
- *o: Tweets with neutral sentiment*
- *t : Total number of Tweets*

then Subjectivity is:

$$Subjectivity = \frac{p+n}{o} = \frac{p+n}{t-(p+n)} \quad (1)$$

and Positivity to Negativity (PN) Ratio is:

$$PNRatio = \frac{p}{n} \quad (2)$$

In Topsy pro the equivalent value is a normalized ratio (0 - 100%) between the positive tweets and tweets with opinions

$$Sentiment = \frac{p}{p+n} \quad (3)$$

D. Seasonal Weighting of Tweets

Season weight was calculated as the given quarter's proportion of the last calendar year. For example, the season weight for calendar Q3.2013 was calculated as below:

$$\frac{\text{Q3.2013 iPhone sales}}{(\text{Q3.2013} + \text{Q2.2013} + \text{Q1.2013} + \text{Q4.2012})}$$

$$= \frac{\text{33.8 million iPhone sales}}{(33.80 + 31.24 + 37.43 + 47.79)}$$

$$= 0.225$$

This proportion number 0.225 is then divided with 0.25 (0.225 / 0.25 = 0.90) to yield the season weight for that particular quarter. So the season weight for Q3.2013 is 0.90 which is multiplied with the 38.72 million tweets for that quarter.

Calculating season weights this way, always 4 quarters back in time, ensures that the calculation is always a mix of Q1, Q2, Q3 & Q4. So only one season weight has to be estimated, which is the latest number for prediction for next quarter. We also tried with 2 years average on the season weighting calculation, but best correlation between iPhone tweets and iPhone sales was obtained with calculation of season weight for 1 year of sales data. The season weighting method with best correlation is based on 1 year of sales data, so an estimated season weight must always go 1 year back. It might be critiqued that once the model get the season weight, it gives the model a strong hint on the number of sales. We do not agree this criticism as most sales prediction models incorporates season weights, as sales fluctuates with considerable season variation. Our use is not much different from the use of season weights in other prediction models.

E. Overall Model

We have made both a linear regression, and a multiple regression prediction model, based on Twitter data. Our final choice was to include the sentiment data from Topsy Pro[2] as our second variable as the sentiment variable improved the correlation and accuracy of the prediction model. Input for the prediction model was then:

$$y = \beta_a \cdot A_{tw} + \beta_p \cdot P_{tw} + \alpha + \varepsilon \quad (4)$$

[1] https://pro.topsy.com/

[2] https://pro.topsy.com

Where

- A_{tw}: *Time lagged and season weighted Twitter data*
- P_{tw}: *Sentiment of* A_{tw}
- *y: iPhone sales in Units*

After using multiple regression analysis in SAS statistical software, we could calculate difference between predicted sales and actual sales, which ended up with 5-10% average error. This concludes our methodological discussion and we now present and discuss the results.

V. Results

As mentioned earlier, we used Topsy Pro, to analyze over 400 million tweets in the period of the Third Quarter of 2007 to the Fourth Quarter of 2013 (Q3.2007- Q4.2013). As Apple publishes iPhone sales by quarters, it became natural to build a prediction model that worked quarterly. A monthly sales prediction model would involve the same principles but our model building followed the structure of quarterly sales data.

Over the period Q3.2007 – Q4.2013 there has been a natural development in the size of the population that is active on Twitter. The development in Twitter users from 2010-2013 could have affected our prediction model. However, from a statistical standpoint, Twitter users showed the same usage patterns during 2010-2013 when tweeting about the iPhone. We did leave out 2007-2009 from our model building for the main reason was a weak link between tweets and sales. 2009 was just atypical in many ways, a statistical outlier – and would have worked as noise for our regression model. From 2010 onwards it is the period of iPhone 3GS, 4, 4S, 5, 5C & 5S.

There is a strong and documented correlation between tweets and iPhone sales in the 2010-2013 period with the Rsquare coefficient of 0.95 and 0.96 for multiple regression with sentiment as the second variable. Output from SAS statistical program is available in the Appendix. Multiple regression analysis – year for year – is a straightforward and quite easy process. However, modeling on a quarterly basis, is a different matter. Only the introduction of seasonal weighting could make our regression model work on a quarterly basis. We have observed that many other prediction models like Morgan Stanley's "Alphawise Smartphone tracker" also use seasonal weighting. We did not copy the principle of seasonal weighting from others, but based on our practical model building professional experience, we realized the necessity of quarterly seasonal weighting. The principles for monthly weighting, would follow – more or less – the same principles if monthly sales data is available. We ended with a prediction model, which showed an average error on app 5-10% for most of the time periods with iPhone sales. The 5-10% average error is close to the average error of the leading predicting methods from Morgan Stanley and IDC – and our model is much simpler and uses less factors (discussion forthcoming). With more research into our model, we expect to get the average error even further down. Figure 2, below presents predicted vs. actual iPhone sales.

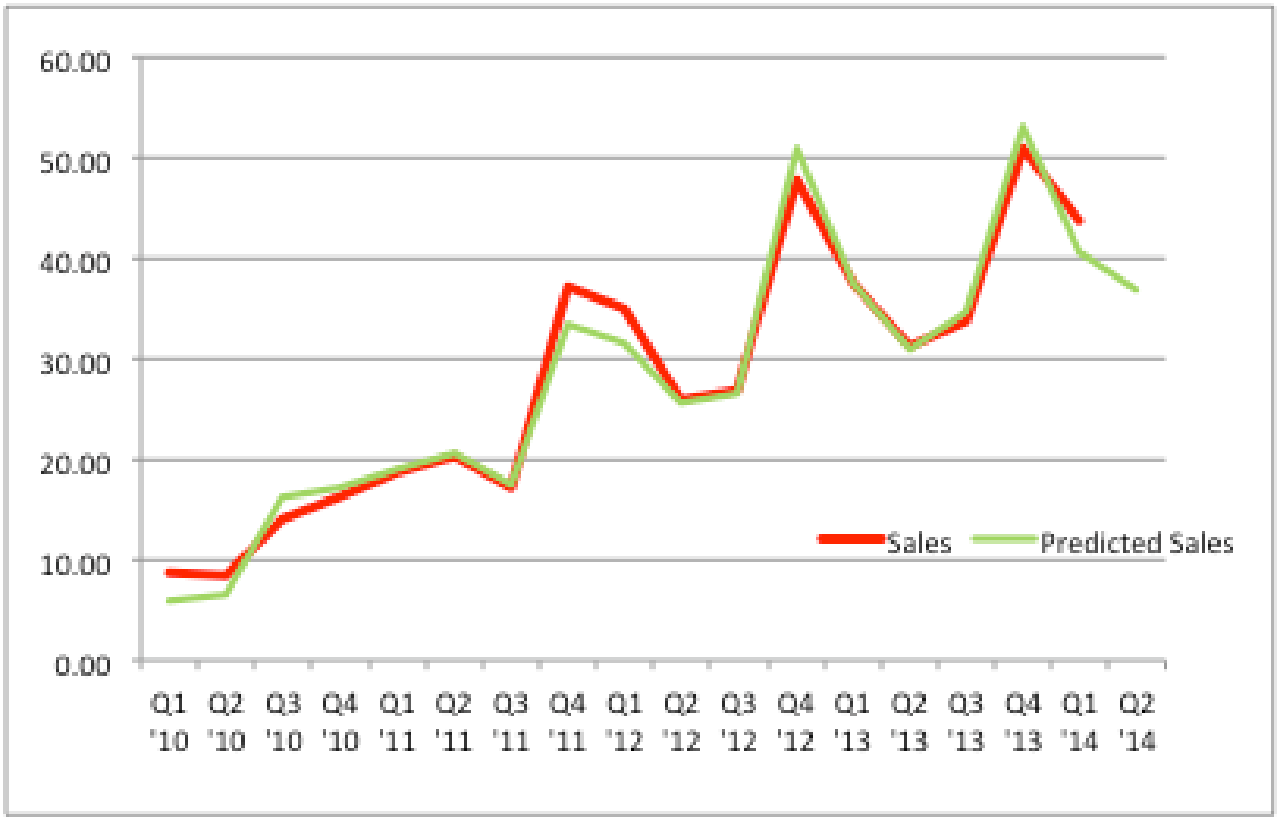


Figure 2 Predicted Quarterly Sales vs. Quarterly Sales

Our main finding is the strength of Twitter as a social data source for predicting smartphone sales. We assume the principles of our prediction model can be used on other products that generate customer opinions and feelings on Twitter. Figure 2 presents the model with final data and shows a prediction of 37 million iPhone sales for Q2.14.

Figure 3 shows that the subjectivity has a declining tendency over time suggesting that people are not as opinionated (passionate) about iPhones as they used to be. This is consistent with the fact that the latest versions of iPhone have not gained any major technological innovations but has shifted from "better" to "more" as in more CPU power, pixel density, and memory. There is a spike in 2011 Q4 around the introduction of iPhone 4S. Also many other black touch sensitive HD screen smart phones with similar capabilities and competitive prices have been introduced on the market since 2010. As the smartphones have increasingly become a mass market product, the "cool" factor of the iPhone has diminished.

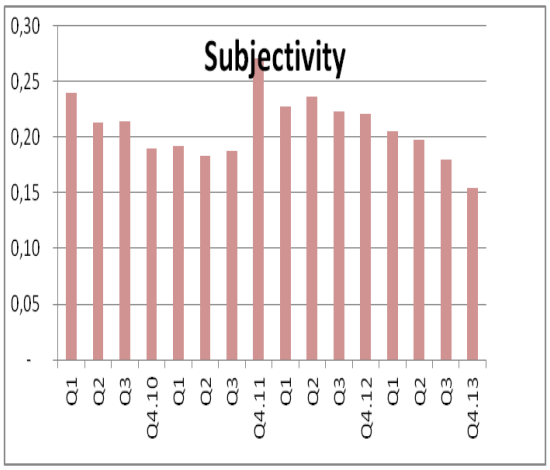


Figure 3 Subjectivity values based on formula (2)

Both the PNRatio shown in Figure 4 and the Sentiment ratio shown in Figure 5 shows a declining tendency that indicates that people are still positive about iPhones but with the overall tendency is decreasing over time. This is consistent with the subjectivity findings as people are less opinionated and less positive about iPhones than before.

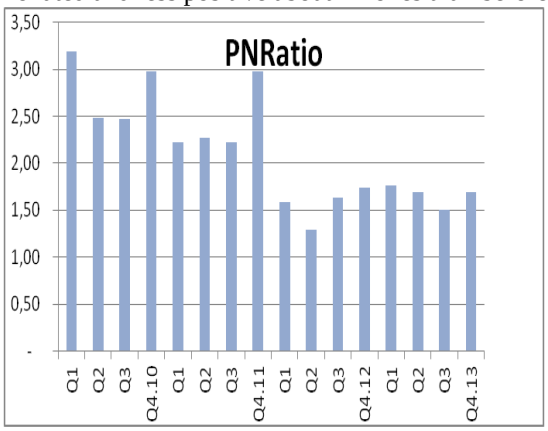


Figure 4 PNRatio values based on formula (3)

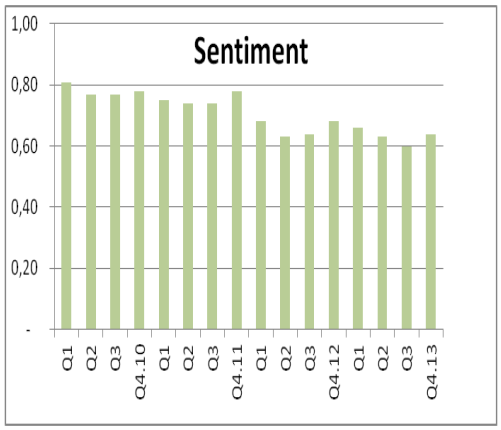
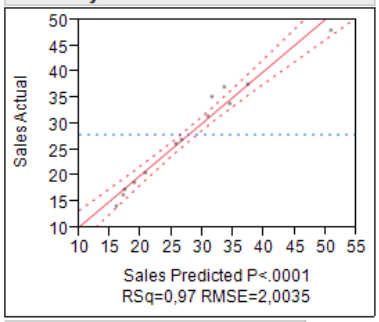


Figure 5 Sentiment values based on formula (4)

Figure 6 presents the output from the statistical software, SAS.

Whole Model

Actual by Predicted Plot



Summary of Fit

RSquare	0,968093
RSquare Adj	0,961712
Root Mean Square Error	2,003467
Mean of Response	27,82308
Observations (or Sum Wgts)	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1217,8657	608,933	151,7067
Error	10	40,1388	4,014	Prob > F
C. Total	12	1258,0045		<,0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22,845247	7,515307	3,04	0,0125*
WtweetRun	0,6987228	0,046536	15,01	<,0001*
Sentiment	-0,210626	0,098453	-2,14	0,0581

Residual by Predicted Plot

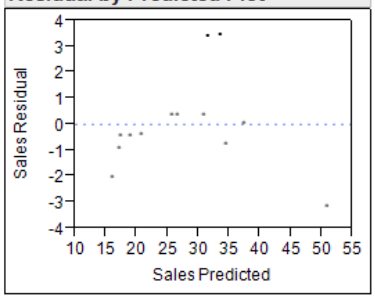


Figure 6 SAS Output for the Prediction Model

VI. DISCUSSION

To summarize, we used Topsy Pro, to analyze over 400 million tweets in the period of the Third Quarter of 2007 to the Fourth Quarter of 2013 (Q3.2007- Q4.2013). We have made both a linear regression, and a multiple regression prediction model, based on Twitter data. Previous research has explored the differences between tweets, retweets and replies on Twitter [31, 32]. However, for our initial model building, we used all the tweets about the iPhone with no differentiation between tweets, retweets, and replies and also with no sentiment analysis.

We treated all the tweets as equal and built the first model. Trying to model with 1, 2 and 3 of the types of tweets, retweets & replies, it became obvious that modeling on all types of tweets (tweets, retweets & replies) gave the best correlation between twitter activities and iPhone sales. One of the metrics of evaluating the impact of tweets and the engagement of followers is called exposure. The exposure of a tweets is calculated as the total potential impressions it has, that is the sum of all followers including each retweet and the sum of their followers and so on. This gives an estimation of the maximum possible users that had the opportunity to read the tweet. It does not remove overlap in users, is simple to calculate and gives a relative performance count to track twitter trends. As a proxy for attention we have chosen only to count original tweets, retweets and replies since these represent active measurable involvement of users.

Social data (like all data) suffers from seasonal variations and therefore requires a cautious approach to extracting the underlying trend. Likewise with sales, for example, smartphone's are a typical Christmas present and have a boosted sales in Q4. To follow the domain-specific theoretical models of AIDA and HoE models, we considered time lagging Twitter data from the beginning. When building the prediction model, we learned that quarter to quarter correlation between Twitter data and iPhone sales did not have the best correlation. We could improve this correlation substantially by pushing Twitter data back in time. We tried many combinations, with 3-6 months of Twitter data, as basis for quarterly sales. For our model building, we chose to weigh all quarterly Twitter data after season weights. Season weights were calculated as the quarter´s sales proportion of a full year. The quarter up for prediction, calendar Q4.2013, got a season weight as a 2 years average, of the season weights in Q4.2012 & Q4.2011. From Adstock models, and other related sales prediction models based on AIDA, we know there is a timelag from customer attention to the actual product purchase. We therefore tested on Twitter data, timelagged back in time – in relation to the quarter we tried to predict. We tried many timelags back in time, and ended up with best correlation between iPhone tweets and iphone sales, for Twitter data pushed back 20 days. An example with predicting calendar Q4.2013: Topsy Pro extract of Tweets containing the phrase "Iphone" and belonging sentiment number, for the period 10 sep 2013 - 10 dec 2013 – which is the basis for predicting calendar period Q4.2013 (1 oct. 2013 – 31 dec 2013). So, the prediction model only predicts quarter sales 20 days before the quarter ends. And 50 days before Apple releases the sales figures.

Our final choice for the model-building was to include the sentiment data from Topsy Pro[3] as our second variable as the sentiment variable improved the correlation and accuracy of the prediction model. Regarding the quality of tweets, the sentiment numbers corresponding to given 3-month period of Twitter data was calculated automatically by the sentiment algorithm of Topsy Pro. As such, the sentiment analysis method is a black box. It is described the algorithm is optimized for English text, and for our 400 million tweets, the majority is English text. For the non-English tweets? In practice, the sentiment numbers improved the correlation between iPhone twitter data and iPhone sales. So we conclude that the Topsy Pro sentiment algorithm also works on non-English text, but presumably with a lower accuracy than on English text.

Our final model is then:

$$y = \beta_a \cdot A_{tw} + \beta_p \cdot P_{tw} + \alpha + \varepsilon \quad (4)$$

Where

- A_{tw}: *Time lagged and season weighted Twitter data*
- P_{tw}: *Sentiment of* A_{tw}
- *y: iPhone sales in Units*

We model the relationship between iPhone sales and iPhone tweets in the period of 2010-2013 and exclude the period of 2007-2009. We find the data for time period of 2007-2010 to be noisy. But from 2010 – 2013 the statistical association is relatively stable, and gives an excellent correlation. Potential reasons could be historical development of user base on Twitter, and also development of the socio-cultural practices of using twitter. We observed a 5-10% average error from our prediction model in formula (1) with the actual sales data over a 2 year period 2012-2013. This average error is not far from the predictions of Morgan Stanley and IDC. For benchmarking purposes, we have identified a few leading prediction methods.

- Morgan Stanley´s "Alphawise Smartphone tracker" by Katy Huberty based on Google trend data, seasonal weighting, and socio economic factors[4].
- IDC's *Worldwide Quarterly Mobile Phone Tracker®, uses* bottom-up methodology[5]
- Steve Milunovich at UBS[6]

[3] https://pro.topsy.com
[4] http://tech.fortune.cnn.com/tag/alphawise/
[5] http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37
[6] http://www.forbes.com/sites/chuckjones/2013/12/03/ubs-analyst- milunovich-upgrades-apple-to-buy-with-650-price-target/

- Peter Misek at Jefferies[7]

Generated By	Information Categories	Typical Investment Debates	Typical Applications
Businesses	Operating Locations	Do a company's operating locations offer a strategic advantage over its competitors	Emerging markets growth; Company competitiveness
	Product Availability	How is the company positioned to meet demand?	Supply-chain bottlenecks; Demand Estimates
	Product Pricing	Is the company able to maintain its prices vis-à-vis its competitors?	Company/Sector Margin Pressure; Inflation; Inventory Growth
	Company Hiring	What positions is the company hiring for?	New product expectations; Growth expectations
Consumers & Clients	Demographics	What is the relative demand from different regions?	Performance of new vs. existing stores/regions
	Product Interest	How successful would a new product launch be? Demand trends	New Product demand Sector demand Consumer spending
	Brand Interest	How is a company's market share evolving	Market share changes

Table 1 Source: Morgan Stanley Research, AlphaWise

None of the corporate market research analysts reveal the technical background for their prediction methods. One of the best predictions comes from Alphawise Smartphone tracker and we shortly compare it to Huberman's model [1]. There is nothing public about the math in this model but there is public description in general terms some of the methodology behind the AlphaWise approach[8]. The generic AlphaWise model is very complex as it takes a vast number of factors into consideration. The factors consist of both Business factors such as Location, Availability, Pricing, and Hiring and Customer related aspects such as Demographics, Product Interest and Brand interest as shown in Table 1. Which of the factors are actually included in the Smartphone tracker application is unclear and it is a qualified guess that Morgan Stanley uses multiple regression.

We did not choose to analyze Samsung Galaxy smartphone sales as "Galaxy" is a common phrase and will create problems when analyzing it on Twitter. On the other hand, the iPhone is a unique smartphone name and is one of the most tweeted products. These were the main criteria for our selection of the iPhone, as a case for a Twitter prediction model. We believe that such technical matters will increasingly become important factors in how companies choose product names. Uniqueness of the product name and hence a possibility for conducting social media analytics will be a point of consideration in the future. This applies for prediction models, customer insights, and many other analytical disciplines that deal with social data.

Regarding generalization, we believe that our approach does generalize to other products of predictions for future years. Different products will require different season weights but building the prediction model of two different products will follow the same principles, with two different set of season weights. The time lag can also be different from product to product. For example, some products could be best predicted with 5 months of Twitter data. Ultimately, the prediction of sales from social data depends on how that specific product´s consumer psychological decision-making process is mirrored on social media channels such as twitter and facebook. Some products will have strong correlation between product posts on social media and product sales in retail and web shops, and some will show weak correlation.

We did consider System Dynamics mathematics, as a model. System dynamics was created during the mid-1950s by Professor Jay Forrester of the Massachusetts Institute of Technology based on a dynamic complex set of differential equations, and causal data relationships. One of the authors of this article have used System Dynamics to predict Christmas tree export from Denmark to Germany. System dynamics is more optimal for complex data pictures containing significant production cycles. It would be possible to build a system dynamics prediction model also

[7] http://www.forbes.com/sites/chuckjones/2013/09/13/jefferies-peter-misek-says-terrible-yields-on-iphone-fingerprint-sensor-hurting-production/

[8] http://tinyurl.com/q2bkxcd

containing twitter data, to predict smartphone sales. A System Dynamics prediction model for smartphone sales could be a natural sequel to this article.

We chose not to experiment with Facebook data for our model building, based on the fact that many product pages on Facebook have about 1% user activity – so for the prediction of smartphone sales, we thought that Facebook was too weak a data source. However, emerging research results are reporting strong correlations between quarterly sales and facebook interactions such as posting, commenting, liking, and sharing [33, 34]. That said, for more in-depth analysis of the smartphone sales, one could include big data analysis of social data from Facebook and other leading social media channels such as Tencent in China. A clear advantage of predicting sales with twitter data is the real-time access to data through Topsy Pro and other analytical tools. Changes in trends and the market can be identified with almost no delays. There is no requirement of phone interviews and traditional observations of customer behavior in this social media analytical approach.

A. *Implications for Organizations*

Our research results have several direct and indirect implications for organizations. The direct implications, obviously, are that sales can be predicted from social media datasets. The indirect implications are that organizations should strategically engage, analyze, and manage social media platforms and mobile applications given the strong correlations between real-world sales and digital-world activities such as social media interactions. An informed and intelligent organizational use of social media to generate competitive advantages [35] requires not only a the adoption of use of technological artefacts for creating valuable affordances [36] for users/consumers but also an understanding of the psychological aspects of how and why consumers share their experiences, interactions, and opinions about products and services as facebook posts, Instagram pictures and tweets [37].

As stated earlier, w believe that the principles of our prediction model can be used on other products that generate customer opinions and feelings on Twitter. In our opinion, big social data analytics that is informed by domain-specific models and theories such as the AIDA (Attention, Interest, Desire, and Action) and the HoE (Heirarchy of Effects) models can yield descriptive, prescriptive, and predictive insights. On that note, we think that the novelty and contribution of our work is in the fact that we conduct theory based big social data analytics (in our case, marketing theories of AIDA and HoE). We believe that this is a small but substantial step towards generating causal explanations and not being limited to documenting statistically significant correlations of sales and social media interactions.

VII. CONCLUSION

Drawing from the theoretical framework of AIDA and Hierarchy of Effects models in advertising combined with an assumptions that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product, we demonstrated how social media data from twitter can be used to predict the sales of iPhones. We developed and evaluated a linear regression model that transforms iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks. This strong correlation between iPhone tweets and iPhone sales becomes marginally stronger after incorporating sentiments of tweets. We discuss our results in terms of a leading industry research as well as academic research based predictive models.

REFERENCES

[1] Asur, S., and Huberman, B.A.: 'Predicting the future with social media', in Editor (Ed.)^(Eds.): 'Book Predicting the future with social media' (IEEE, 2010, edn.), pp. 492-499

[2] Bakshy, E., Simmons, M.P., Huffaker, D., Teng, C., and Adamic, L.: 'The social dynamics of economic activity in a virtual world', ICWSM2010. http://misc.si.umich.edu/publications/18, 2010

[3] Bollen, J., and Mao, H.: 'Twitter mood as a stock market predictor', Computer, 2011, pp. 91-94

[4] Dorr, D.H., and Denton, A.M.: 'Establishing relationships among patterns in stock market data', Data & Knowledge Engineering, 2009, 68, (3), pp. 318-337

[5] Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R.: 'Mining the stock market (extended abstract): which measure is best?', in Editor (Ed.)^(Eds.): 'Book Mining the stock market (extended abstract): which measure is best?' (ACM, 2000, edn.), pp. 487-496

[6] Kharratzadeh, M., and Coates, M.: 'Weblog Analysis for Predicting Correlations in Stock Price Evolutions', in Editor (Ed.)^(Eds.): 'Book Weblog Analysis for Predicting Correlations in Stock Price Evolutions' (2012, edn.), pp.

[7] Mittermayer, M.-A.: 'Forecasting intraday stock price trends with text mining techniques', in Editor (Ed.)^(Eds.): 'Book Forecasting intraday stock price trends with text mining techniques' (IEEE, 2004, edn.), pp. 10 pp.

[8] Huber, J., Landherr, A., Probst, F., and Reisser, C.: 'Stimulating User Activity On Company Fan Pages In Online Social Networks', ECIS 2012 Proceedings. Paper 188. http://aisel.aisnet.org/ecis2012/188, 2012

[9] Lin, Z., and Goh, K.Y.: 'Measuring the Business Value of Online Social Media Content for Marketers', ICIS 2011 Proceedings. Paper 16. http://aisel.aisnet.org/icis2011/proceedings/knowledge/16 2011

[10] Zimbra, D., Fu, T., and Li, X.: 'Assessing public opinions through Web 2.0: a case study on Wal-Mart', ICIS 2009 Proceedings. Paper 67. http://aisel.aisnet.org/icis2009/67, 2009

[11] Heath, D., Singh, R., Ganesh, J., and Kroll-Smith, S.: 'Exploring Strategic Organizational Engagement in Social Media: A Revelatory Case', ICIS 2013 Proceedings. http://aisel.aisnet.org/icis2013/proceedings/EBusiness/13/, 2013

[12] Larson, K., and Watson, R.T.: 'The value of social media: toward measuring social media strategies', in Editor (Ed.)^(Eds.):

‘Book The value of social media: toward measuring social media strategies’ (2011, edn.), pp.
[13] Dinter, B., and Lorenz, A.: ‘Social Business Intelligence: a Literature Review and Research Agenda’, in Editor (Ed.)^(Eds.): ‘Book Social Business Intelligence: a Literature Review and Research Agenda’ (2012, edn.), pp.
[14] Rosemann, M., Eggert, M., Voigt, M., and Beverungen, D.: ‘Leveraging social network data for analytical CRM strategies: the introduction of social BI’, in Editor (Ed.)^(Eds.): ‘Book Leveraging social network data for analytical CRM strategies: the introduction of social BI’ (AIS Electronic Library (AISeL), 2012, edn.), pp.
[15] Geva, T., Oestreicher-Singer, G., Efron, N., and Shimshoni, Y.: ‘Do Customers Speak Their Minds? Using Forums and Search for Predicting Sales’, Available at SSRN: http://ssrn.com/abstract=2294609 or http://dx.doi.org/10.2139/ssrn.2294609, 2013
[16] Koppius, O.: ‘The Value of Online Product Buzz in Sales Forecasting’, ICIS 2010 Proceedings. Paper 171. http://aisel.aisnet.org/icis2010_submissions/171, 2010
[17] Oh, C., and Sheng, O.: ‘Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement’, in Editor (Ed.)^(Eds.): ‘Book Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement’ (2011, edn.), pp.
[18] Seebach, C., Pahlke, I., and Beck, R.: ‘Tracking the Digital Footprints of Customers: How Firms can Improve Their Sensing Abilities to Achieve Business Agility’, ECIS 2011 Proceedings, 2011
[19] Wu, L., and Brynjolfsson, E.: ‘The future of prediction: how Google searches foreshadow housing prices and quantities’, ICIS 2009 Proceedings. Paper 147. http://aisel.aisnet.org/icis2009/147, 2009
[20] Zhang, W., and Lau, R.: ‘The Design of a Network-Based Model for Business Performance Prediction’, ICIS 2013 Proceedings. http://aisel.aisnet.org/icis2013/proceedings/KnowledgeManagement/10/, 2013
[21] vd Reijden, P., and Koppius, O.R.: ‘The Value of Online Product Buzz in Sales Forecasting’, in Editor (Ed.)^(Eds.): ‘Book The Value of Online Product Buzz in Sales Forecasting’ (2010, edn.), pp. 171
[22] Nann, S., Krauss, J., and Schoder, D.: ‘Predictive Analytics On Public Data-The Case Of Stock Markets’, roceedings of the 21st European Conference on Information System. http://www.staff.science.uu.nl/~Vlaan107/ecis/files/ECIS2013-0615-paper.pdf, 2013
[23] Li, H., and Leckenby, J.: ‘Examining the Effectiveness of Internet Advertising Formats’, in Schumann, D., and Thorson, E. (Eds.): ‘Internet Advertising: Theory and Research’ (Lawrence Erlbaum Associates, 2007), pp. 203-224
[24] Kelman, H.C.: ‘Compliance, identification, and internalization: Three processes of attitude change’, The Journal of conflict resolution, 1958, 2, (1), pp. 51-60
[25] Cialdini, R., and Goldstein, N.: ‘Social influence: Compliance and conformity’, Annual Review of Psychology, 2004, 55, (1), pp. 591-621
[26] Schumann, D., and Thorson, E.: ‘Internet Advertising: Theory and Research’ (Lawrence Erlbaum Associates, 2007. 2007)
[27] Lavidge, R.J., and Steiner, G.A.: ‘A model for predictive measurements of advertising effectiveness’, The Journal of Marketing, 1961, pp. 59-62
[28] Barry, T.E.: ‘The development of the hierarchy of effects: An historical perspective’, Current issues and Research in Advertising, 1987, 10, (1-2), pp. 251-295
[29] Ray, M.: ‘Marketing communication and the hierarchy of effects’, New models for communication research, 1973, pp. 146-175
[30] Belch, G.E., Belch, M.A., Kerr, G.F., and Powell, I.: ‘Advertising and promotion: An integrated marketing communications perspective’ (Mcgraw-Hill, 2008. 2008)
[31] Romero, D., Galuba, W., Asur, S., and Huberman, B.: ‘Influence and passivity in social media’, Machine Learning and Knowledge Discovery in Databases, 2011, pp. 18-33
[32] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P.: ‘Measuring user influence in twitter: The million follower fallacy’, ICWSM, 10, 2010, pp. 10-17
[33] Mukkamala, R., Hussain, A., and Vatrapu, R.: ‘Towards a Formal Model of Social Data’, IT University Technical Report Series, 2013, TR-2013-169, pp. https://pure.itu.dk/ws/files/54477234/ITU_TR_54472013_54477169.pdf
[34] Mukkamala, R., Hussain, A., and Vatrapu, R.: ‘Towards a Set Theoretical Approach to Big Data Analytics’, Proceedings of IEEE Big Data 2014, Anchorage, USA, in press/2014
[35] Vatrapu, R.: ‘Understanding Social Business’, in Akhilesh, K.B. (Ed.): ‘Emerging Dimensions of Technology Management’ (Springer, 2013), pp. 147-158
[36] Vatrapu, R.: ‘Explaining culture: an outline of a theory of socio-technical interactions’, Proceedings of the 3rd ACM International Conference on Intercultural Collaboration (ICIC 2010), 2010, pp. 111-120
[37] Kunst, K., and Vatrapu, R.: ‘Towards A Theory Of Socially Shared Consumption: Literature Review, Taxonomy And Research Agenda’, Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, 2014, pp. ISBN 978-970-9915567-9915560-9915560