Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data

Niels Buus Lassen, Ravi Vatrapu, Lisbeth la Cour, Rene Madsen, and Abid Hussain

Article in proceedings (Final published version)

CITE: Towards a Theory of Social Data : Predictive Analytics in the Era of Big Social Data. / Buus Lassen, Niels; Vatrapu, Ravi; la Cour, Lisbeth; Madsen, Rene; Hussain, Abid. Symposium i anvendt statistik: 25.-27. januar 2016. ed. / Peter Linde. København : Danmarks Statistik, 2016. p. 241-256.

Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data

Niels Buus Lassen, Ravi Vatrapu, Lisbeth la Cour, René Madsen, Abid Hussain, Copenhagen Business School

INTRODUCTION

In this chapter, we will advance a theory of social data that distinguishes between constituent dimensions of social graph (i.e., socio-technical affordances of social media networks) and those of social text (i.e., communicative and linguistic properties of social media interactions) as distinct but complementary elements of predictive big social data analytics. Additionally, to illustrate the validity and applicability of our proposed theory, we adhered to the schematic steps advocated by Shmueli and Koppius (2011) in building empirical predictive models that blend social graph analysis with social text analysis to: (1) compute correlations between social data from multiple social media platforms (i.e., Facebook and Twitter) and the financial performance (i.e., quarterly revenues) of corporate entities (i.e., iPhones and H&M), as well as; (2) make predictions about the future performance of these corporate entities. In doing so, we endeavor to provide an answer to the following research question: *How can big social data analytics be utilized to predict business performance?*

This paper comprises four sections, inclusive of this introduction.. In Section 2, we construct our theory of social data by extending Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Section 3 outlines our methodological strategy for extracting and analyzing big social data to build empirical predictive models of business performance. Results from analyzing these empirical predictive models are also reported in Section 3. The last section, Section 4, summarizes the: (1) implications of this study to both theory and practice; (2) insights to be gleaned towards informing the application of predictive analytics to big social data; (3) possible limitations in the interpretation of our empirical findings, and; (4) probable avenues for future research.

TOWARDS A THEORY OF SOCIAL DATA

To bridge the knowledge gaps in extent literature, we advance a theory of social data that extends Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Social media (e.g., Facebook and Twitter), at the highest level of abstraction, involve social entities interacting with: (a) technologies (e.g., an individual using the Facebook app on his/her smartphone), and; (b) other social entities (e.g., the same individual liking a picture of a friend on the Facebook app). Vatrapu (2008, 2010) labelled

these interactions as *sociotechnical interactions* (see also Vatrapu and Suthers 2010). Socio-technical interactions yield electronic trace data that we termed as ***social data***. To derive a theory for social data, we must first determine the constituents of socio-technical interactions. As acknowledged by Vatrapu (2010), socio-technical interactions are realized through: (a) a social entity's perception and appropriation of *socio-technical affordances*, as well as; (b) the structures and functions of *technological intersubjectivity* (Vatrapu 2010).

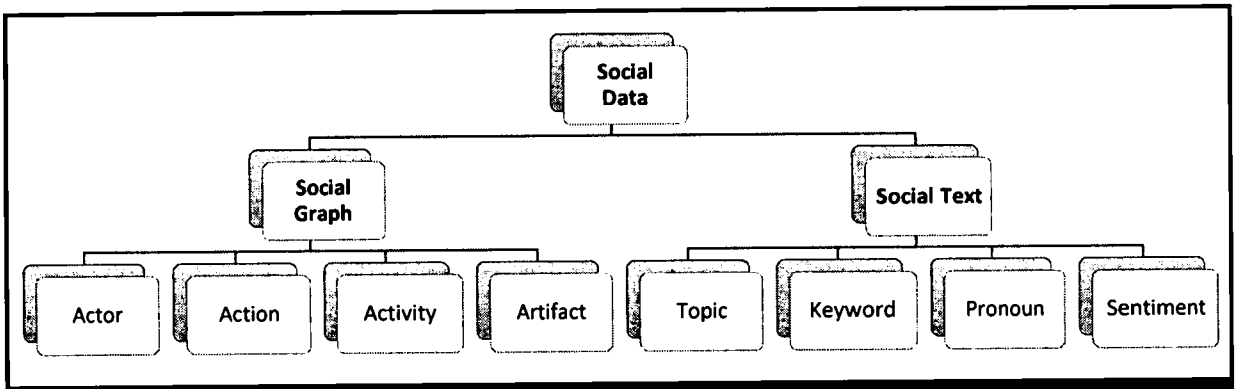

Figure 1: Theory of Social Data

As an illustration of our theory, consider the earlier example of an individual liking a friend's picture on the Facebook app. The performance of such a simple sociotechnical interaction already activates multiple social data elements: an *actor* (i.e., individual) performing an *action* (i.e., liking) on an *artifact* (i.e., Facebook app) for the purpose of expressing a *sentiment* (i.e., like) and contributing to a collective *activity* (i.e., expanding the social network timeline). Such micro social-technical interactions, when amassed in large volumes, constitute the macro world of big social data, the core premise of this paper.

METHODOLOGY AND ANALYTICAL FINDINGS

In this section, we presents details about the collection, preparation, exploration, selection, modelling and reporting of two big social data sets to illustrate different aspects of our proposed theory of social data. In general, we adhered to the methodological schematic recommended by Shmueli and Koppius (2011, p. 563) for building empirical predictive models. The remainder of this section is organized in accordance with Shmueli and Koppius's (2011) eight methodological steps of predictive model building as depicted in Figure 2.

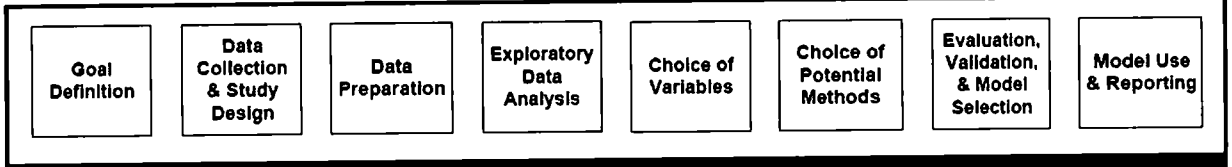

Figure 2: Methodological Steps in Predictive Model Building [Shmueli and Koppius 2011, p. 563]

Step 1: Goal Definition

Our primary goal was to build empirical predictive models of sales from big social data. More specifically, by applying predictive analytics to big social data, we strive to model and accurately predict the real-world numerical outcomes of quarterly sales of Apple iPhone & H&M revenues.

Step 2: Data Collection and Study Design

We discuss the rationale for the study design first followed by details on data collection.

Study Design: The study was designed to collect and analyze big social data sets that serve as illustrative case studies for predictive analytics. Therefore, we deliberately introduce variance into both the predicted variable of sales as well as the predictor variables of social data attributes.

With regard to the predicted variable of sales, we sought to incorporate variance in terms of *product types* (i.e., Apple iPhone: consumer electronics and H&M: fashion-clothes) and *sales channels* (i.e., offline and online; direct and retail). As for the predictor variables of big social data, we incorporated variance in terms of *social media platforms* (i.e., Facebook and Twitter), *theory of social data attributes* (Social Graph: actors, actions, artifacts and Social Text: keywords and sentiment), *dataset sizes* (few millions to hundreds of millions of data points), and *data time periods* (few months to years). Table 1 summarizes the characteristics of the two big social datasets that have been collected, processed and analyzed in this paper.

Table 1: Big Social Datasets Collected for Predictive Analytics

Company	Data Source	Time Period	Size of Dataset	Mapping to Social Data Attributes
Apple[1]	Twitter	2007 → June, 2015	500 million+ tweets containing "iPhone"	▪ **Social Text**: Keyword ("iPhone") ▪ **Social Text**: Sentiment
H & M[2]	Facebook	January 01, 2009 → March, 2015	~15 million Facebook events	▪ **Social Graph**: Actions (Total Likes) ▪ **Social Graph**: Artifacts (Posts and Comments) ▪ **Social Graph**: Actors (H&M + Non-H&M) ▪ **Social Text**: Sentiment

Data Collection: We now present details on the methods and tools used for data collection for the two big social datasets.

Twitter (Apple: "iPhone")

We collected over 500 million tweets containing the phrase "iPhone" in the period 2007-2015 (till March, 2015) via Topsy Pro Analytics[3]. Technically, our data collection did not connect to the Twitter firehose, but rely on a Twitter API solution with full access to all Twitter data.

[1] URL: https://www.apple.com.
[2] URL: http://www.hm.com.
[3] URL: https://pro.topsy.com.

Facebook (H&M)

Facebook wall data was extracted by a specialized big social data analytics tool called SODATO. SODATO[4] is an IT artifact, a software solution that is custom built for collecting, storing, processing, and analyzing big social data from social media platforms. The construction of SODATO is not only informed by our proposed theory of social data, but it is also methodologically built in adherence to Sein et al.'s (2011) Action Design Research (ADR) principles. Technically, SODATO utilizes the APIs provided by the social network vendors (e.g., Facebook open source API named as Graph API). Table 2 gives an overview of the social data collected by SODATO from the official Facebook walls of H&M.

Table 2: Overview of Facebook Data

Company	Official Facebook Wall: Name (id)	Time Period	Facebook Posts	Facebook Comments	Facebook Likes
H&M	Hm (21415640912)	January, 2009 → March, 2015	127,920	366,863	14,367,067

Sales (Apple and H&M)

Data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. This concludes the presentation of the methods and tools used for data collection and overviews of the different big social datasets. We now discuss the third step in predictive analytics prescribed by Shmueli and Koppius (2011), data preparation.

Step 3: Data Preparation

Twitter (Apple: "iPhone")

We searched for the keyword "iPhone" in Topsy Pro, which then returned number of all tweets (i.e., Tweets, retweets, and replies) for the time period specified, and with sentiment numbers pre-calculated. These numbers form the basis for our 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 modelling.

Facebook (H&M)

Facebook data was first fetched by SODATO via the Facebook Graph API and was then pre-processed and aggregated in order to make it available on demand for Analytics engine and at the end to the visualization module. The grouping of different analysis units was done in accordance

[4] URL: http://cssl.cbs.dk/software/sodato.

with the different attributes of the theory of social data (Social Graph: actors, actions and artefacts and Social Text: sentiment).

Sales (Apple and H&M)

As mentioned earlier, data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. These were tabulated into Excel spreadsheets together with quarterly measures of social graph and social text.

Step 4: Exploratory Data Analysis

Shmueli and Koppius (2011) stated that during exploratory data analysis: "each question, rather than each construct, would be treated as an individual predictor. In addition to exploring each variable, examining the correlation table between BI and all of the predictors would help identify strong predictor candidates and information overlap between predictors (candidates for dimension reduction)" (p. 657).

Our objectives for the explorative data analytics were twofold: First, to build on the seminal regression model of Asur and Huberman (2010) for predicting movie revenues from twitter sales. Second, based on the Hierarchy of Effects (HoE) (Lavidge and Steiner 1961) and the AIDA (Attention, Desire, Interest, and Action) (Li and Leckenby 2007) domain-specific models of advertising and sales respectively, to explore different predictor variables, different data transformations of the predictor variables in terms of time lagging and different options for seasonal weighting of the predicted variable, sales.

We organize this section in the order of the two datasets (Apple iPhone tweets & H&M facebook) and describe the explorative data analysis conducted on the respective big social data sets that had already been collected and prepared.

"iPhone" Dataset

For the iPhone dataset, we selected the social data attributes of *social graph: actions* (tweets, re-tweets, replies and mentions) and *social text: keyword* ("iphone") and *social text: sentiments* (positive, negative, and neutral). We explored the temporal dynamics of the social data measures *social graph: actions* and *social text: sentiments* for the filter social text: keyword ("iPhone") directly on the Topsy Pro web site. We then explored the dataset by creating two predictor variables: quantity of tweets and quality of tweets as described below.

Quantity of Tweets

To provide an example, for the time period of September 10, 2013 to December 10, 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.

Time Lagging of Tweets

As mentioned earlier, our predictive analytics method is informed by both the theory of social data and the AIDA and HoE domain-specific models. The key analytical challenge in social data predictive analytics is to model real-world outcomes from social data measures of social graph (actions, artefacts, activities and actions) and social text (topics, keywords, pronouns and sentiments). From the AIDA and the HoE domain-specific models and based on standard industry practice, we explored different options for time-lagging of social data measures as proxy for the sales funnel inherent in the time period between a potential customer becoming aware of the product, developing an interest in the product, having a desire for it and ultimately deciding to obtain it typically by a sales transaction. We experimented with different time-lags and found 20 days to be the statistically optimal value for the iPhone twitter dataset. As will be discussed later, we found different time lags for different datasets. It is important to note that even though the AIDA and HoE models can help in the exploration of the time lag in the first place and a partial explanation of its existence, they do not theoretically predict a particular value. This, we hope will be addressed with research advances in computational social science in general and predictive analytics in particular

Seasonal Weighting of Sales

Again, based on the AIDA and HoE models, and given the product life cycle of new models and new operating system releases of Apple iPhone, we conducted season weighting of the quarterly sales. Seasonal weights were 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:

$$\frac{\text{Q3.2013 iPhone Sales}}{(\text{Q3.2013} + \text{Q2.2013} + \text{Q1.2013} + \text{Q4.2012})} = \frac{\text{33.8 million iPhone Sales}}{(33.80 + 31.24 + 37.43 + 47.79)} = \underline{0.225}$$

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. An estimated season weight for prediction must always go 1 year back. Next, we present the exploratory data analysis of the H&M dataset.

H&M Dataset: Following Shmueli and Koppius (2011)'s advice for exploratory data analysis step of predictive analytics, we explored the predictive power of several different variables constructed from the theory of social data. In summary, we created two categories of the social data attribute of *social graph: a*ctors (H&M and Non-H&M). We then calculated the distribution of the social data attribute of *social graph: artefacts* (posts, comments, and likes) across the two

actor types. With respect to the social data attribute of social text: sentiments (positive, negative, and neutral), based on the sentiment analysis of the social text artefacts (posts and comments) discussed earlier, we calculated distributions of sentiments across different kinds of artifacts and actors (i.e. positive sentiments on posts by H&M actors (wall administrators), positive sentiments on posts by Non-H&M actors etc.). We then calculated the quarterly aggregates of these different measures of social data attributes and evaluated the statistical correlation with respect to quarterly sales. Surprisingly, statistically significant positive correlations with quarterly revenues were observed for negative sentiments on total posts.

Logarithmic Transformation and Time lagging of Facebook Likes

Informed by the correlational analysis above and based on further exploratory data analysis with different predictor variables, we selected the logarithmic transformation of 40 days' time lagged total likes per quarter as the main predictor variable from the array of social data attributes listed in Tables 4 and 5 above.

Seasonal Weighting of Quarterly Sales

As with iPhone quarterly sales, we used a weighted measure of the quarterly revenues of H&M to account for seasonal variation of sales corresponding to fashion cycles (i.e., Fall, Winter, Spring and Summer Collections) and holidays across the different H&M markets.

Step 5: Choice of Variables

Choice of the predictor variables is based on careful considerations of theory, domain-specific knowledge and empirical association with predicted variables (Shmueli and Koppius, 2011). Based on exploratory data analysis, the following variables were chosen for the two big social datasets as summarized in Table 3.

Table 3: List of Chosen Predictor Variables

Company / Product	Time Period of Quarter	Seasonal Weighting of Dependent Variable [Sales]	Independent Variable #1 (including info on transformation)	Independent Variable #2	Time-Lagging of Independent Variable #1	Time-Lagging of Independent Variable #2
iPhone Sales (Quarterly)	Calendar Quarters	+	No of tweets over 3 months period	sentiment	20 days	20 days
H&M	Quarter ends 1 month before calendar quarter: Q4.2014 is from September 01 → November 30	+	LOG (No of total likes over 3 months period)	none	40 days	none

Step 6: Choice of Methods

As discussed earlier, we analytical objective was to not only build on but also extend the predictive modelling of Asur and Huberman (2010). As such, we chose regression modelling as the method and sought to extend the method by using time lagged and transformed predictor variables of social data measures and seasonally adjusted predicted variables.

Step 7: Evaluation, Validation and Model Selection

Our overall predictive analytics model for big social data analytics is stated below:

$$y = \beta_a \times A_t + \beta_p \times P_t + \beta_d \times D + \varepsilon$$

Where:

$A_t = \sum A_{st}$

A_{st} = Social media activity in terms of actions by actors on artifacts associated with sales at time t (Social Graph Attributes)

A_t = Accumulated time-lagged social media activity associated with sales at time t

P_t = Polarity at time t (Social Text Attribute)

D = Distribution factor (Sales Channel Attribute)

We now present the specific prediction models for the two big social datasets of iPhone & H&M.

Social Data Predictive Analytics Model for iPhone Sales

We modelled the relationship between iPhone sales and iPhone tweets in the period of 2010-2014 and excluded the period of 2007–2009. While the data for time period of 2007–2009 is noisy, the statistical association is relatively stable for 2010–2013 and gives an excellent correlation. Potential reasons could be historical growth of user base on Twitter, and also the development of socio-cultural practices of using twitter. The predictive model for iPhone sales is:

Predicted Sales of iPhones Sold (in millions) = WtweetRun * 0,6987228 + Sentiment * (-0,210626) + 22,845247 (intercept)

where

- WtweetRun is the season weighted tweets count for 3 months period time lagged by 20 days back from the sales quarter
- Sentiment is the sentiment for tweets for 3 month period time lagged by 20 days back the from sales quarter

Figure 3 presents the statistical output for the iPhone predictive model.

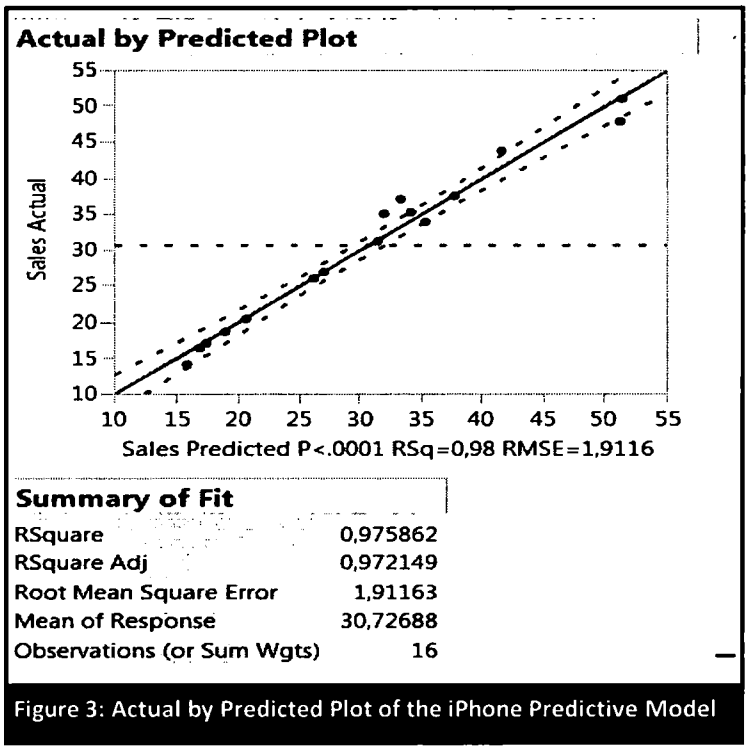

Figure 3: Actual by Predicted Plot of the iPhone Predictive Model

Figure 4 depicts the graph for the iPhone predictive model.

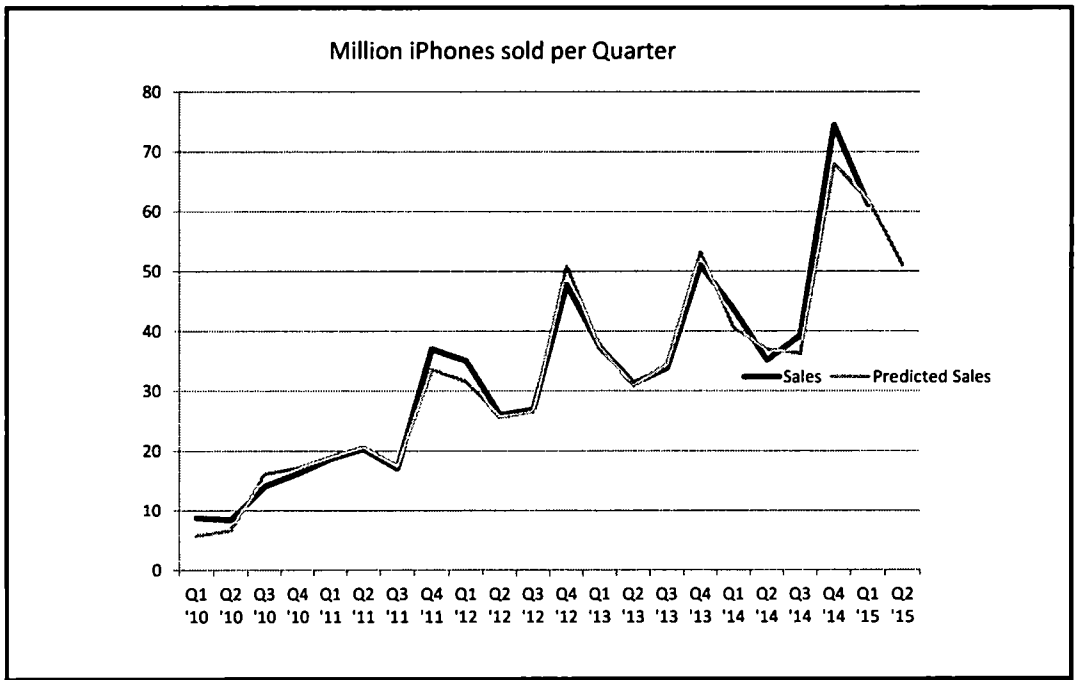

Figure 4: Predicted vs, Actual Sales of iPhones

Social Data Predictive Analytics Model for H&M Revenues

Based on the linear regression for the period 2011-2014, our predictive analytics model for 2014 is given by the following equation:
Predicted Revenue for H&M (in billions SEK) = 2,28 billion SEK * seasonweight * LOG (Facebook total likes time lagged by 40 days back over a 3 months period) + 5,45 billion SEK (the intercept)

Figure 5 presents the SAS output for the 2011-2014 predictive modelling

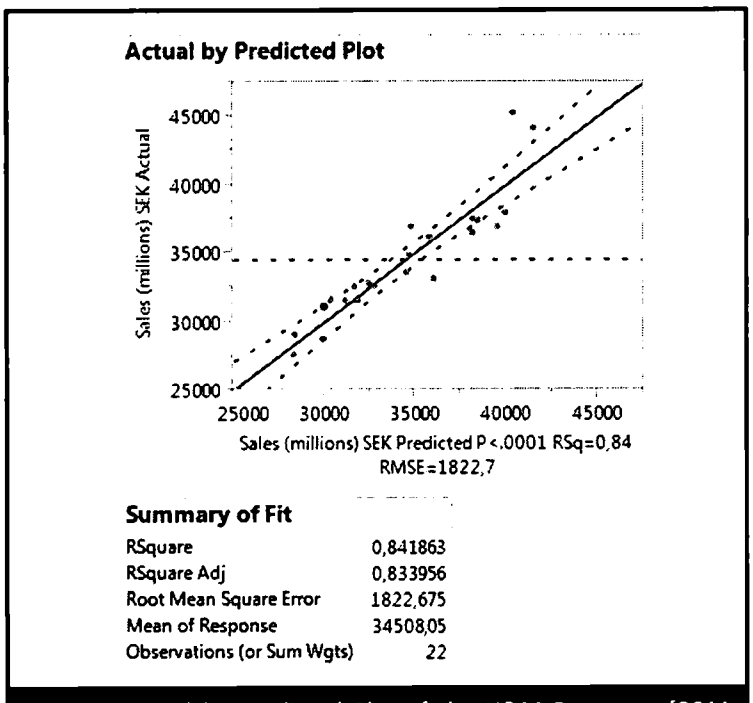

Figure 5: Actual by Predicted Plot of the H&M Revenues [2011-2014]

However, for the period of 2010-2013, based on the linear regression of data for 2009-2013, the predictive model is:

Predicted Revenue for H&M (in billions) = 1,67 billion SEK * seasonweight * LOG (facebook total likes time lagged by 40 days back over a 3 months period) + 13 billion SEK (the intercept)

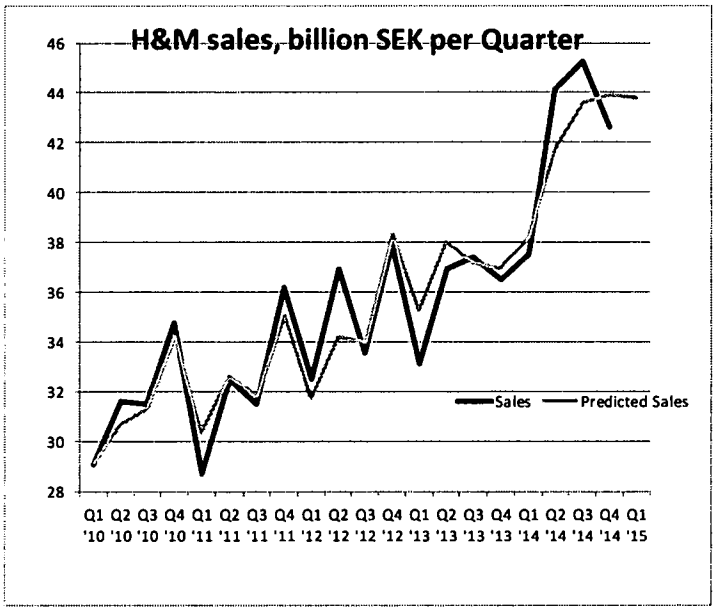

Figure 6: Predicted vs. Actual Revenues of H&M

Figure 6 depicts the combined chart of predicted vs. actual revenues of H&M

Step 8: Model Use and Reporting

In this step, we focus on predictive accuracy and meaning (Shmueli & Koppius, 2011). With regard to our prediction models, we observed a 5-10% average error from our prediction model with the actual sales data over 3 year period 2012-2014. In the case of iPhone, 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 for iPhone sales.

- Morgan Stanley's "Alphawise Smartphone tracker" by Katy Huberty based on Google trend data, seasonal weighting, and socio economic factors[5].
- IDC's *Worldwide Quarterly Mobile Phone Tracker®, uses* bottom-up methodology[6]
- Steve Milunovich at UBS[7]

DISCUSSION

[5] URL: http://www.forbes.com/sites/chuckjones/2014/03/19/morgan-stanleys-alphawise-smartphone-tracker-has-iphone-demand-ahead-of-consensus.
[6] URL: http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37.
[7] URL: http://www.forbes.com/sites/chuckjones/2013/12/03/ubs-analyst-milunovich-upgrades-apple-to-buy-with-650-price-target.

Though predictive analytics has been touted to be a major growth segment for research into social media, there is only a handful of studies to-date that have managed to capitalize on this opportunity. This paper thus takes a small but concrete step towards furthering this research agenda by advancing and validating a theory of social data for enhancing predictive analytics. Detailed implications for theory and practice are elaborated below.

Implications for Theory

This paper makes a novel contribution to extant literature on several fronts. First, past studies on social networks have typically progressed as two separate research streams with one seeking to comprehend the structural properties of such networks (i.e., social network analysis) (e.g., Johnson et al. 2014; Moser et al. 2013; Putzke et al. 2010; Shi et al. 2014; Trier 2008; Trier and Richter 2014; Whelan 2007; Whelan et al. 2013) and the other trying to infer value from the communicative content shared within these networks (i.e., sentiment analysis) (e.g., Cheung et al. 2012; Clemons et al. 2006; Jensen et al. 2013; Li and Hitt 2010; Mudambi and Schuff 2010). Yet, at the same time, there is evidence to suggest that invaluable insights could be gleaned from research that considers the structural properties and communicative content of social networks in tandem (see Butler et al. 2014; Chau and Xu 2012; Füller et al. 2014; Gasson and Waters 2013; Gray et al. 2011; Moser et al. 2013; Trier and Richter 2014). Therefore, in distinguishing between social graph and social text as constituent elements of social data, our proposed theory gives equal prominence to the two aforementioned research streams by embracing the structural properties and communicative content of social media.

Second, our theory of social data is the first to bring clarity to plausible dimensions that could be incorporated into empirical predictive models for social media (see Figure 1). By deriving constituent dimensions of social graph (i.e., actor, action, activity and artifact) and social text (i.e., topic, keywords, pronoun and sentiment), we enlarge the pool of options for applying predictive analytics to big social data. Third, we demonstrate the applicability of our proposed theory through the construction of empirical predictive models that are invariant to the kind of social media platform (i.e., Facebook and Twitter) from which data is extracted and the type of corporate entities (i.e., financial performance of H&M and iPhone) to be predicted, be it companies or products. In this sense, our proposed theory of social data can be deemed as a cornerstone for future studies of predictive big social data analytics to build upon.

Last but not least, beyond predictive analytics, we believe that our proposed theory of social data can also aid in the generation of holistic frameworks for computational social science in general and big social data analytics in particular. So far, computational methods, formal models and software tools for big social data analytics have been largely confined to graph theoretical approaches (Gross and Yellen 2005) in the likes of social network analysis (Borgatti et al. 2009),

which in turn is informed by the social philosophical approach of relational sociology (Emirbayer 1997). As far as we know, there is no other unified modeling approaches to social data that assimilates conceptual, formal, software, analytical and empirical domains (Mukkamala et al. 2013). Recent work (e.g., Vatrapu et al. 2014a, 2014b) has sought to outline an alternative approach to the predominant triad of relational sociology, graph theory and social network analysis, which are founded on associational sociology (Latour 2005), set theory and fuzzy set theory (Ragin 2000) as well as social set analysis (Mukkamala et al. 2014).

Implications for Practice

This paper should be of interest to practitioners for three reasons. First, our empirical results bear direct and indirect implications for companies. Naturally, a direct and obvious implication from this study is the proof that business performance can be predicted from big social data. By extracting and analyzing data from multiple social media platforms (i.e., Facebook and Twitter) to predict the financial performance of both companies (i.e., H&M) and products (i.e., iPhone), we are able to show that the predictive power of big social data is neither constrained by the social media platform nor the type of parameter to be predicted. For this reason, the indirect implications are that companies should proactively engage and strategically manage social media platforms in order to benefit from the strong correlations between social media interactions and sales performance. Second, by delineating social data into elements of social graph and social text, we provide companies with a schema of the elements to pay attention to on social media platforms. In order for companies to generate competitive advantage from social media, they must not only recognize the structural relationships within social networks, they must also value the opinions and sentiments embodied within social media content. Finally, this study is the first of its kind to take into account the existence of a time-lag from the moment a potential customer becomes aware of a product to the instance he/she decides to acquire it via a sales transaction when building empirical predictive models. In a way, this study highlights the importance of social media as an inexpensive forum for companies to continuously maintain product awareness in the minds of consumers.

Limitations

There are several limitations to the work reported here. First, we lack multiple cases to extensively evaluate and validate the overall prediction model. A second limitation is the emerging challenge for predictive analytics from social data associated with increasing sales in emerging markets such as China with its own unique social media ecosystem. By and large, the social media ecosystem of China does not overlap with that of Western countries to which Facebook and Twitter belong. We suspect that the effect of non-overlapping social media ecosystems might be somewhat ameliorated for Veblen goods such as iPhones given the

conspicuous consumption aspirations of a global middle class. This however remains an analytical challenge and restricts the predictive power of our H&M prediction model. A third limitation of the paper is that the theory of social data is limited to a cross-sectional framework of social data in terms of social graph (i.e., actors, actions, activities and artefacts) and social text (i.e., topics, keywords, pronouns and sentiments). As such, our theory of social data does not outline a process model, which might be more pertinent to predictive analytics. A fourth limitation arises from the representativeness of social media data. That said, as far as predictive analytics of real-world activities is concerned, social media datasets might be adequately representative as long as the basic premise of a social media action being a proxy for a user's attention to that particular real-world activity holds true. Our theory of social data will only cease to be valid if and when a user's social media action (such as a tweet about an "iPhone") is not a proxy for that user's attention towards the "iPhone" object. In our view, this fundamental disjunction between social media actions and real-world attention is the Achilles's Heel of predictive analytics with social data and might partially explain the spectacular drop in accuracy for once popular prediction models like the Google Flu Prediction System. A fifth and final limitation of our study, as far as our knowledge goes, is the lack of theoretical explanation for the empirical values for the time lags both in the nominal sense and the relative sense of divergence between Facebook and Twitter.

Future Work

For future work, we envision several projects that could spawn from this research as outlined below.

Going beyond the traditional and pre-dominant sentiment classification of social text and towards domain-specific classifiers such as AIDA and HoE for predicting sales. This will require not only sophisticated computational linguistics methods and tools but also critical contributions from domain experts (e.g., for training datasets in the case of supervised machine learning algorithms).

Investigating other predictor variables such as socio-economic factors, confidence, trust, loyalty etc. Essentially. Moving towards "thick models" of human users and narrowing the social media user and real-world consumer gap for non-digital products and services.

Combining social media data with other online sources such as Google Trends or in-house data of enterprise systems such as ERP and CRM.

REFERENCES

Asur, S. and Huberman, B. A. "Predicting the Future with Social Media," in Proceedings of the *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT) (Vol. 1), 2010, pp. 492-499.

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G. "Network Analysis in the Social Sciences," *Science* (323:5916), 2009, pp. 892-895.

Butler, B. S., Bateman, P. J., Gray, P. H. and Diamant, E. I. "An Attraction–Selection–Attrition Theory of Online Community Size and Resilience," *MIS Quarterly* (38:3), 2014, pp. 699-728.

Chau, M. and Xu, J. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), 2012, pp. 1189-1216.

Cheung, M. Y., Sia, C. L. and Kuan, K. K. "Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an ELM Perspective," *Journal of the Association for Information Systems* (13:8), 2012, pp. 618-635.

Clemons, E. K., Gao, G. G. and Hitt, L. M. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), 2006, pp. 149-171.

Emirbayer, M. "Manifesto for a Relational Sociology," *The American Journal of Sociology* (103:2), 1997, pp. 281-317.

Füller, J., Hutter, K., Hautz, J. and Matzler, K. "User Roles and Contributions in Innovation-Contest Communities," *Journal of Management Information Systems* (31:1), 2014, pp. 273-308.

Gasson, S. and Waters, J. "Using A Grounded Theory Approach to Study Online Collaboration Behaviors," *European Journal of Information Systems* (22:1), 2013, pp. 95-118.

Gray, P. H., Parise, S. and Iyer, B. "Innovation Impacts of Using Social Bookmarking Systems," *MIS Quarterly* (35:3), 2011, pp. 629-643.

Gross, J. L. and Yellen, J. *Graph Theory and Its Applications*, CRC press, 2005.

Hussain, A. and Vatrapu, R. "Social Data Analytics Tool," DESRIST 2014, *Lecture Notes in Computer Science* (LNCS), 8463(Springer), 2014, pp. 368–372.

Jensen, M. L., Averbeck, J. M., Zhang, Z. and Wright, K. B. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective," *Journal of Management Information Systems* (30:1), 2013, pp. 293-324.

Johnson, S. L., Faraj, S. and Kudaravalli, S. "Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment," *MIS Quarterly* (38:3), 2014, pp. 795-808.

Lassen, N., Madsen, R. and Vatrapu, R. "Predicting iPhone Sales from iPhone Tweets," in Proceedings of the *18th IEEE Enterprise Computing Conference* (EDOC 2014), Ulm, Germany, 2014.

Latour, Bruno (2005). *Reassembling the social an introduction to actor-network-theory*. Oxford New York:

Oxford University Press. ISBN 9780199256044.

Lavidge, R. J. and Steiner, G. A. "A Model for Predictive Measurements of Advertising Effectiveness," *Journal of Marketing* (25:6), 1961, pp. 59-62.

Li, H. and Leckenby, J. "Examining the Effectiveness of Internet Advertising Formats," in D. Schumann & E. Thorson (eds.), *Internet Advertising: Theory and Research*, Lawrence Erlbaum Associates, 2007, pp. 203-224.

Li, X. and Hitt, L. M. "Price Effects In Online Product Reviews: An Analytical Model And Empirical Analysis," *MIS Quarterly* (34:4), 2010, pp. 809-831.

Moser, C., Ganley, D. and Groenewegen, P. "Communicative Genres as Organizing Structures in Online

Communities–of Team Players and Storytellers," *Information Systems Journal* (23:6), 2013, pp. 551-567.

Mudambi, S. M. and Schuff, D. "What Makes A Helpful Online Review? A Study Of Customer Reviews on Amazon.Com," *MIS Quarterly* (34:1), 2010, pp. 185-200.

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Formal Model of Social Data," *IT University Technical Report Series*, TR-2013-169, 2013. [Available online at: https://pure.itu.dk/ws/files/54477234/ITU_TR_54472013_54477169.pdf, accessed October 14, 2014]

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Set Theoretical Approach to Big Data Analytics," in *Proceedings of IEEE Big Data 2014*, Anchorage, United States of America, 2014.

Putzke, J., Fischbach, K., Schoder, D. and Gloor, P. A. "The Evolution Of Interaction Networks In Massively Multiplayer Online Games. *Journal of the Association for Information Systems* (11:2), 2010, pp. 69-94.

Ragin, C. C. *Fuzzy-Set Social Science*, University of Chicago Press, 2000.

Sein, M., Henfridsson, O., Purao, S., Rossi, M. and Lindgren, R. "Action Design Research," *MIS Quarterly* (35:1), 2011, pp. 37-56.

Shi, Z., Rui, H. and Whinston, A. B. "Content Sharing in A Social Broadcasting Environment: Evidence From Twitter," *MIS Quarterly* (38:1), 2014, pp. 123-142.

Shmueli, G. and Koppius, O. R. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), 2011, pp. 553-572.

Trier, M. "Research Note-Towards Dynamic Visualization For Understanding Evolution Of Digital Communication Networks," *Information Systems Research* (19:3), 2008, pp. 335-350.

Trier, M. and Richter, A. "The Deep Structure of Organizational Online Networking–An Actor□Oriented Case Study," *Information Systems Journal* (Advance copy) 2014.

Vatrapu, R. "Understanding Social Business," In K. B. Akhilesh (ed.), *Emerging Dimensions of Technology Management*, New Delhi: Springer, 2013, pp. 147-158.

Vatrapu, R. and Suthers, D. "Intra-and Inter-Cultural Usability in Computer-Supported Collaboration," *Journal of Usability Studies* (5:4), 2010, pp. 172-197.

Vatrapu, R. Cultural Considerations in Computer Supported Collaborative Learning," *Research and Practice in Technology Enhanced Learning* (3:2), 2008, pp. 159-201.

Vatrapu, R. K. "Explaining Culture: An Outline of a Theory of Socio-Technical Interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, Copenhagen, Denmark, 2010, pp. 111-120.

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings," in *Computational Social Science: Contagion, Collective Behaviour, and Networks*, Oxford: University of Oxford, 2014a, pp. 22-24. [available online at: http://cssworkshop.oii.ox.ac.uk]

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "Towards a Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Empirical Findings," in *Proceedings of the 5th Annual Social Media & Society International Conference 2014*, Toronto, Canada, 2014b.

Whelan, E. "Exploring Knowledge Exchange In Electronic Networks of Practice," *Journal of Information Technology* (22:1), 2007, pp. 5-12.

Whelan, E., Golden, W. and Donnellan, B. "Digitizing The R&D Social Network: Revisiting The Technological Gatekeeper," *Information Systems Journal* (23:3), 2013, pp. 197-218.