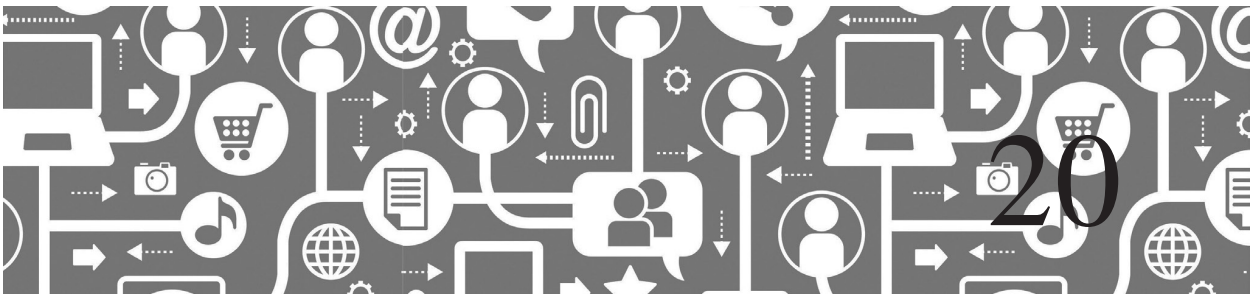

20

Predictive Analytics with Social Media Data

Niels Buus Lassen, Lisbeth la Cour, and Ravi Vatrapu

This chapter provides an overview of the extant literature on predictive analytics with social media data. First, we discuss the difference between predictive vs. explanatory models and the scientific purposes for and advantages of predictive models. Second, we present and discuss the foundational statistical issues in predictive modelling in general with an emphasis on social media data. Third, we present a selection of papers on predictive analytics with social media data and categorize them based on the application domain, social media platform (Facebook, Twitter, etc.), independent and dependent variables involved, and the statistical methods and techniques employed. Fourth and last, we offer some reflections on predictive analytics with social media data.

INTRODUCTION

Social media has evolved into a vital constituent of many human activities. We increasingly share several aspects of our private, interpersonal, social, and professional 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 by employing a variety of research methods as documented in this handbook (Quan-Haase & Sloan, Chapter 1, this volume). Social data analytics is not only informing, but also transforming existing practices in politics, marketing, investing, product development, entertainment, and news media. This chapter focuses on predictive analytics with social media data. In other words, how social media data has been used to predict processes and outcomes in the real world.

Recent research in the field of Computational Social Science (Cioffi-Revilla, 2013; Conte et al., 2012; Lazer et al., 2009) has shown how data resulting from the widespread adoption and use of social media channels such as Facebook and Twitter can be used to predict outcomes such as Hollywood

movie revenues (Asur & Huberman, 2010), Apple iPhone sales (Lassen, Madsen, & Vatrapu, 2014), seasonal moods (Golder & Macy, 2011), and epidemic outbreaks (Chunara, Andrews, & Brownstein, 2012). Underlying assumptions for this research stream on predictive analytics with social media data (Evangelos et al., 2013) 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 (Vatrapu et al., 2015).

PREDICTIVE MODELS VS. EXPLANATORY MODELS

At the outset, we find that the difference between predictive and explanatory models needs to be emphasized. Predictive analytics entail the application of data mining, machine learning and statistical modelling to arrive at *predictive models* of future observations as well as suitable methods for ascertaining the *predictive power* of these models in practice (Shmueli & Koppius, 2011). Consequently, predictive analytics differ from explanatory models in that the latter aims to: (1) draw statistical inferences from validating causal hypotheses about relationships among variables of interest, and; (2) assess the explanatory power of causal models underlying these relationships (Shmueli, 2010). This crucial distinction between explanatory and predictive models is best surmised by Shmueli & Koppius (2011) in the following statement: "whereas explanatory statistical models are based on underlying *causal relationships between theoretical constructs*, predictive models rely on *associations between measurable variables*" (p. 556). For example, in political science, explanatory models have investigated the extent to which social media platforms such as Facebook can function as online public spheres (Robertson & Vatrapu, 2010; Vatrapu, Robertson, & Dissanayake, 2008) in terms of users' interactions and sentiments (Hussain, Vatrapu, Hardt, & Jaffari, 2014; Robertson, Vatrapu, & Medina, 2010a,b). On the other hand, predictive models in political science sought to predict election outcomes from social media data (Chung & Mustafaraj, 2011; Sang & Bos, 2012; Skoric, Poor, Achananuparp, Lim, & Jiang, 2012; Tsakalidis, Papadopoulos, Cristea, & Kompatsiaris, 2015).

Distinguishing between explanation and prediction as discrete modelling goals, Shmueli & Koppius (2011) argued that any model, which strives to embrace both explanation and prediction, will have to trade-off between explanatory and predictive power. More specifically, Shmueli & Koppius (2011) claim that predictive analytics can advance scientific research in six scenarios: (1) generating new theory for fast-changing environments which yield rich datasets about difficult-to-hypothesize relationships and unmeasured-before concepts; (2) developing alternate measures for constructs; (3) comparing competing theories via tests of predictive accuracy; (4) augmenting contemporary explanatory models through capturing complex patterns which underlie relationships among key concepts; (5) establishing research relevance by evaluating the discrepancy between theory and practice; and (6) quantifying the predictability of measureable phenomena.

This chapter discusses predictive modelling of (big) social media data in social sciences. The focus will be entirely on what is often referred to as predictive models: models that use statistical and/or mathematical modelling to predict a phenomenon of interest. Furthermore, the focus will be on prediction in the sense of forecasting a future outcome of the phenomenon of interest as such predictions are the ones that have so far received most attention in the literature. To illustrate the concepts, models, methods and evaluation of results we use examples from economics and finance. The general principles are, however, easily employed to other social science fields as well, for example,

marketing. The concepts and principles that this section discusses are of a general nature and are informed by Hyndman & Athanasopoulos (2014) and Chatfield (2002).

This chapter does not discuss applicable software solutions. However, it is worth mentioning that there exist quite a few software packages with more or less automatic search procedures when it comes to model specification. A few ones are, for example, SAS, SPSS and the Autometrics package of OxMetrics.

PREDICTIVE MODELLING OF SOCIAL MEDIA DATA

When performing predictive analysis on social media data researchers often have to make a lot of decisions along the way. Examples of the most important decisions or choices will be discussed in the sections below.

The phenomenon of interest and the type of forecasts

Quite often the focus will be on a single outcome (univariate modelling – one model equation) where the goal is to derive a prediction or forecast of, for example, sales in a company or the stock price of the company. In some cases, more than one outcome will be of interest and then a multivariate approach in which more than one relationship or model equation is specified, estimated, and used at the same time is worth considering. From now on let us assume that the phenomenon of interest is sales of a company and the social media data are among the factors that are considered as explanatory for the outcome. The discussion will then relate to the univariate case. At this stage, a decision is also necessary in relation to the data frequency. Is the predictive model supposed to be applied to forecast monthly sale, quarterly sales or sales of an even higher frequency like weekly or daily?

The data

Once the phenomenon of interest is identified, decisions concerning the data to be used have to be made. Data can be of different types: time series (e.g. sales per month or sales per day), cross sectional (e.g. individuals such as customers, for a given period in time) or longitudinal/panel (a combination of the former two such as a set of customers observed through several months). Predictive models can be relevant for all these types of data and many of the basic principles for analysis are quite similar. In the remaining parts of this section, for simplicity the focus will be on time series only.

As social media data have been growing in volume and importance during the last 10 years, in some cases the final number of observations for modelling may be rather limited as the dependent variable may reflect accounting and book-keeping and be relatively low-frequency like monthly or quarterly in nature. If this is the case, there may be a limit to how advanced models can be used. In other cases, daily data may be available and more complex models may be considered.

The frequency of the data is also important for model specification itself. With more high frequency data, a researcher may discover more informative dynamic patterns compared to a case with less frequent data. Consider a case where sales of a company need to be forecasted. If the reaction time from increased activity on the Facebook page of the company to changes in sales is short (e.g. just a couple of days) then if sales are available only on a monthly basis the lag pattern between explanatory factors and outcome may be difficult to identify and use.

In many cases there will be a large set of potential explanatory factors that may be included in various tentative model specifications. Social media data may be just a part of such data and it will be important to also include other variables. The quality as well as the quantity of data is very important for building a successful predictive model.

Social media data and pre-processing

When researchers consider using social media data for predictive purposes, at the outset the social media data will be collected at the level of the individual action (e.g. a Facebook 'like' or a tweet) and in order to prepare the data to enter a predictive model some pre-processing will be necessary. Often the data will need to be temporally aggregated to match the temporal aggregation level of the outcome, for example, monthly data. Also as some of the inputs from social media are text variables, some filtering, interpretation, and classification may be necessary. An example of the latter would be the application of a supervised machine learning algorithm that classifies the posts and comments into positive, negative or neutral sentiments (Thelwall, Chapter 32, this volume). At the current moment it is mainly the pre-processing of the social media data that is considered challenging from the computational aspects of big data analytics (Council, 2013). Once the individual actions (posts, likes, etc.) are temporally aggregated and classified, the set of potential explanatory factors are usually rather limited and as the outcome variables are of fairly low frequencies like monthly or quarterly (stock market data are actually sometimes used at a daily frequency) which means that the modelling process deviates less from more classical approaches within predictive modelling.

In search of a model equation – theory-based versus data-driven?

In very general terms a model equation will identify some relationship between the phenomenon of interest (y) and a set of explanatory factors. The relationship will never be perfect either due to un-observable factors, measurement errors or other types of errors.

The general equation: $y = f(\text{explanatory factors}) + \text{error}$

Where f describes some relationship between what is inside the parenthesis and y.

In principle, linear, non-linear, parametric, non-parametric and semi-parametric models may be considered. In general, non-linear models will require more data points/observations than linear models as the structures they search for are more complex.

There is a range of possible starting points for the search process. At one end lies traditional econometrics where the starting point is often an economic or behavioural theory that will guide the researcher in finding a set of potential explanatory factors. At the other end of the range machine learning algorithms will help identify a relationship from a large set of social media data and other potential explanatory factors. The advantage of starting from a theory-based model specification is that the researcher may be more confident that the model is robust in the sense that the identified relationship is reliable at least for some period of time. Without a theory the identified structure may still work for predictions in the short run but may be less robust and in general will not add much to an understanding of the phenomenon at hand. In between pure theoretically inspired models and models based on data pattern discoveries are many models that include elements of both categories. As theoretical models are often more precise when it comes to selection of explanatory factors for the more fundamental or long-run relationships they may be less precise when it comes to a description of dynamics and a combination that allows for a primary theoretically based long-run part may prove more useful.

To finalize the discussion of theory-based versus data-driven model selection the concept of causality is often useful. If a causal relationship exists a change in an explanatory factor is known to imply a change in the outcome. A model that suffers from a lack of a causal relationship suffers from an endogeneity problem (a concept used in econometrics). A model that suffers from an endogeneity problem will not be useful for tests of a

theory of for policy evaluations. If the only purpose of the model is forecasting, identification of a causal relationship is of less importance as a strong association between the explanatory factors and the outcome may be sufficient. However, without causality the predictive model may be considered less robust (more risk of a model break-down) to general changes in structures and society and hence may be best at forecasting in the short run. If this is the case, some sort of monitoring on a continuous basis to identify a model break-down at an early stage is advisable.

Fitting of a predictive model

In this step the researcher will adapt the mathematical specification of the predictive model to the actual data. In the case of a linear regression model this is done by estimation using the ordinary least squares (OLS) method or the maximum likelihood (ML). For non-linear models such as neural networks, some mathematical algorithm is used. In rare cases estimation of a model is not possible (e.g. in case of perfect multicollinearity of a linear regression model). In such a case the researcher has to re-think the model specification.

Estimation (the use of a formula or a procedure) may in itself sound simple, but already at this stage the researcher has to specify the set-up to be used for model evaluation in the following step as they are highly dependent.

Even though it may seem natural to use as many data point as possible for the model fitting, there are other considerations to take into account as well. For the estimation step, it is stressed that in addition to the decision of estimation or fitting method, a decision on exactly which sample or part of the sample to use for estimation is of importance too.

Evaluation of a predictive model for forecasting purposes

The true test of a predictive model that is to be used for forecasting of future values of the outcome of interest is by investigating the out-of-sample properties of the model.

This statement calls for the need of an estimation (or training) sample and an evaluation (or test) sample. As a good in-sample model fit does not ensure good forecasting properties of a predictive model, the evaluation process then naturally starts by an analysis of the in-sample properties of the model and extends to an out-of-sample analysis.

In-sample evaluation of the model

The first thing to note is that if the model has a theoretical foundation the signs of the estimated coefficients will be compared to the signs expected from the theory.

A second thing to be aware of is whether the model fulfils the underlying statistical assumptions (these may differ depending on the type of model in focus). In classical linear regression modelling, problems such as autocorrelation and heteroscedasticity will need attention and a study of potential outliers is of high importance. When forecasting is the final purpose of the model multicollinearity is of less importance. Finally, indicators in relation to the functional form specification may provide useful information on how to improve the model.

The overall fit of the model may be captured by measures such as R^2, adjusted R^2, the family for measures based on absolute or squared errors (e.g. MSE, RMSE, MAE, MAPE), and information criteria such as AIC, and BIS. A small warning is justified here as too much emphasis on obtaining a good fit may result in overfitting of the model which is not necessarily desirable when the purpose of the model is forecasting.

Out-of-sample evaluation

For an out-of-sample evaluation study the model is used to forecast values for a time period that was not used for the estimation of the model. In the 'pure' case neither

future values of the explanatory factors nor future values of the outcome are known and the model that is used to obtain the forecast will need to rely on lagged values of the explanatory factors or to use predicted values of the explanatory factors. In the former case, the specification of the model equation in terms of lags will set a limit to how many periods into the future the model can predict. In many cases an out-of-sample forecast evaluation will rely on sets of one step ahead predictions, but predictions for a longer forecast horizon (e.g. six months ahead for a model specified with monthly data) are also sometimes considered.

Once the out-of-sample forecasts are obtained it is possible to calculate forecast errors and to study their patterns. Focus areas will be of directional nature (the trend in the outcome captured), as they may be related to predictability of turning points and summary measures for the errors will again prove useful (e.g. MSE, MAPE, etc.) but this time for the forecasted period only. The idea of splitting the sample into different parts for evaluation can be extended in various ways using cross-validation (Hyndman & Athanasopoulos, 2014).

Using a predictive model for forecasting purposes

Once a model has been chosen some considerations concerning its implementation are important. This topic is very much related to the overall phenomenon and problem; hence a general discussion is difficult to provide.

There is, however, one type of considerations that deserves mentioning: how often the model needs re-estimation or specification updating. Given that often the general data pattern is quite robust, the specification updating may only take place in case of new variables becoming available or in case a sufficiently large number of data points have become available such that more complex structures could be allowed for.

Finally, from a practical perspective a combination of forecasts from different basic predictive models is also a possibility and quite popular in certain fields.

CATEGORIZED LIST OF PREDICTIVE MODELS WITH SOCIAL MEDIA DATA

Table 20.1 below presents a selected list of research papers on predictive analytics with social media data categorized across different application domains in terms of social media platform (Facebook, Twitter, etc.) and the independent and dependent variables involved. For conceptual exposition and literature review on the predictive power of social media data (see Gayo-Avello et al. (2013)).

Application Domains

As can be seen from Table 20.1, there have been many predictive models of sales based on social media data. Such predictive models work for the brands that can command large amounts of human attention on social media, and therefore generate big data on social media. Examples are iPhone sales, H&M revenues, Nike sales, etc., which are all product categories around which there is a possibility to have large volumes and ranges of opinions on social media platforms. For brands and products that don't generate large volumes of social media data, for instance, insurance, banking, shipping, basic household supplies, etc. the predictive models tend not to work. One explanation for the successful performance of the predictive models is that social media actions can be categorized into the phases of the different domain-specific models from the application domains of marketing, finance, epidemiology, etc. For example, the actual stock price for Apple is in rough terms mainly based on discounted historical sales and expectations to future sales. If social media can model sales, then

Table 20.1 Categorization of Research Publications on Predictive Analytics with Social Media Data

Reference	*Social Data*	*Dependent Variables*	*Independent Variables*	*Statistical Methods*
Asur & Huberman (2010)	Twitter	Movie revenue	Twitter activity, sentiment and theatre distribution	Time-Series Multiple Regression Model
Lassen et al. (2014)	Twitter	iPhone sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Bollen & Mao (2011)	Twitter	Dow Jones Industrial Average	Calm, Alert, Sure, Vital, Kind and Happy	Time-Series Multiple Regression Model
Voortman (2015)	Google Trends	Car sales	Google trend data car names	Time Series Linear Regression Model
Vosen & Schmidt (2011)	Google Trends	Consumer spending	Real personal income y, interest rates on 3-month Treasury Bills I and stock prices s (measured on S&P 500), Google Trend, and consumer spending t-1	ARIMA/Time Series Multiple Regression Model
Choi & Varian (2012)	Google Trends	Sales of cars, homes and travel	Historical sales and Google trend variable	Simple Seasonal AR Models and Fixed-Effects Models
Chung & Mustafaraj (2011)	Twitter	Political election outcome	Twitter collective sentiment	Linear Regression
Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer (2011)	Twitter	Political alignment	Twitter hashtags	SVM trained on hashtag metadata
Bothos, Apostolou, & Mentzas (2010)	IMDB, Flixster, Yahoo Movies, HSX, Twitter, RottenTomatoes.com	Movie Academy Award winners	Measures from IMDB, Flixster, Yahoomovies, HSX, Twitter, RottenTomatoes.com	Multivariate Distribution Models
Culotta (2010)	Twitter	Detecting influenza outbreaks	Twitter keywords	Time-Series Multiple Regression Model
Dijkman, Ipeirotis, Aertsen, & van Helden (2015)	Twitter	Many types of sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Eysenbach (2011)	Twitter	Total number of citations	Twimpact variable (number of tweetations within n days after publication)	Multi-Variate/Linear Regression
Gruhl, Guha, Kumar, Novak, & Tomkins (2005)	Blogs	Sales	Product/brand mentions	Time-Series using Cross-Correlation
Jansen, Zhang, Sobel, & Chowdury (2009)	Twitter	Brand variables	Twitter sentiment variables	Time-Series Linear Regression Models
Li & Cardie (2013)	Twitter	Early stage influenza detection	Twitter texts about flu	Unsupervised Bayesian Model based on Markov Network
Radosavljevic, Grbovic, Djuric, & Bhamidipati (2014)	Tumblr	Sport results and number of goals	Team and player mentions	Poisson Regression Model using Maximum Likelihood Principle
Ritterman, Osborne, & Klein (2009)	Twitter	Stock-Prices	Historical prices, unigrams and bigrams, Twitter activity	SVR Regression using Unigrams and Bigrams

Sang & Bos (2012)	Twitter	Dutch election outcome	Twitter texts and sentiments	Time Series Multiple Regression Model
Shen, Wang, Luo, & Wang (2013)	Twitter	Entity belonging	Twitter texts	Decision tree, KAURI/LINDEN method
Skoric et al. (2012)	Twitter	Election outcome Singapore	Twitter activity	Time Series Linear Regression Model
Tsakalidis et al. (2015)	Twitter	Election outcomes EU	Twitter texts	Linear Regression (LR), Gaussian Process (GP) and Sequential Minimal Optimization for Regression (SMO)
Tumasjan, Sprenger, Sandner, & Welpe (2010)	Twitter	Election outcomes Germany	Twitter texts	Probability Models
Yu, Duan, & Cao (2013)	Google blogs, Boardreader and Twitter compared to Google News	Firm equity value	Variables for activity and sentiment	Time Series Multiple Regression Model
Hughes, Rowe, Batey, & Lee (2012)	Twitter and Facebook	Socialising and info exchange	Big5 personality traits, NFC and sociability	Time-Series Multiple Regression Model
Krauss, Nann, Simon, Gloor, & Fischbach (2008)	Forums	Movie success and academy awards	Intensity, positivity and trendsetter variables	Time-Series Multiple Regression Model
Seiffertt & Wunsch (2008)	Several	Variables on financial markets	Many types discussed	Different Model Types Discussed
Tang & Liu (2010)	Flickr and YouTube	Online behaviors	Social Dimension variables	SocioDim, several advanced models combined
Karabulut (2013)	Facebook	Stock prices	Facebook GNH (General national happiness), positivity, negativity	Time-Series Multiple Regression Model
Mao, Counts, & Bollen (2014)	Twitter	UK, US, and Canadian stock markets	"Bullish" or "bearish" mentions on Twitter	Time-Series Multiple Regression Model
Bollen, Mao, & Zeng (2011)	Twitter	DJIA	Twitter moods and feelings	Self-Organizing Fuzzy Neural Network
Bollen, Mao, & Pepe (2011)	Twitter	Socio-economic events	Twitter moods and feelings	Extended version of: Profile of mood states
Eichstaedt et al. (2015)	Twitter	Heart attacks	Anger, stress and fatigue	Time-Series Multiple Regression Model
De Choudhury, Gamon, Counts, & Horvitz (2013)	Twitter	Depression	Language, emotion, style, ego-network, and user engagement	Support Vector Machine
De Choudhury, Counts, & Horvitz (2013)	Twitter	Postpartum changes in emotion and behaviour	Engagement, emotion, ego-network and linguistic style	Support Vector Machine
De Choudhury, Counts, Horvitz, & Hoff (2014)	Facebook	Postpartum depression	Social activity and interaction	OLS Regression Model
Weeks & Holbert (2013)	Facebook, Twitter, YouTube	Dissemination of News Content in Social Media	Gender, age, web engine news search, email news activity and cell phone activity	Decision Tree Model
Gilbert & Karahalios (2009)	Facebook	Tie strength	15 communication variables	Time-Series Multiple Regression Model
Won et al. (2013)	Weblog social media data	Suicide	Suicide related words and mentions	Time-Series Multiple Regression Model

there is a high potential for the associated stock price to also being modelled with social media data. In the case of epidemiology, all social media texts on flu can also be categorized in to the different domain-specific phases of spread, incubation, immunity, resistance, susceptibility etc.

Social Media Data Types

For modelling stock prices, Twitter and Google Trends have proven to be the best platforms. Twitter and Google Trends beat Facebook for stock price modelling because of higher data volume and immediacy. On the other hand, Facebook data have been successfully used for modelling sales, human emotions, personalities and human relations to a brand. In general, picture and video based social media platforms such as Instagram, YouTube and Netflix are becoming more prevalent and we expect them to become more relevant for predictive models in the future.

Independent and Dependent Variables

As can be seen from Table 20.1, a wide range of dependent variables have been modelled: sales, stock prices, Net Promoter Score, happiness, feelings, personalities, interest areas, social groups, diseases, epidemics, suicide, crime, radicalization, civil unrest. The independent variables used reflect the human social relations to the dependent variables mainly consist of measures of social media activity, feelings, personalities and sentiment.

Statistical Methods Employed

We find that a wide range of statistical models for predictive analytics have been used including Regression, Neural Network, SVM, Decision Trees, ARIMA, Dynamic Systems, Bayesian Networks, and combined models.

In the next section, we present an illustrative case study of predictive modelling with big social data.

AN ILLUSTRATIVE CASE STUDY OF PREDICTIVE MODELLING

In this section, we demonstrate how social media data from Twitter and Facebook can be used to predict the quarterly sales of iPhones and revenues of clothing retailer, H&M, respectively. Based on a conceptual model of social data (Vatrapu, Mukkamala, & Hussain, 2014) consisting of Interactions (actors, actions, activities, and artifacts) and Conversations (topics, keywords, pronouns, and sentiments), and drawing from the domain-specific theories in advertising and sales from marketing (Belch, Belch, Kerr, & Powell, 2008), we developed and evaluated linear regression models that transform (a) iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks (Lassen et al., 2014) and (b) Facebook likes into a prediction of the global revenue of the fast fashion company, H&M. Our basic premise is that social media actions can serve as proxies for user's attention and as such have predictive power. The central research question for this demonstrative case study was: *To what extent can Big Social Data predict real-world outcomes such as sales and revenues?* Table 20.2 below presents the dataset collected for predictive analytics purposes of this case study.

We adhered to the methodological schematic recommended by Shmueli & Koppius (2011) for building empirical predictive models. We built on and extended the predictive analytics method of Asur & Huberman (2010) and examined if the principles for predicting movie revenue with Twitter data can also be used to predict iPhone sales and

Table 20.2 Overview of Dataset

Company	*Data Source*	*Time Period*	*Size of Dataset*
Apple	Twitter	01–2007 to 10–2014	~500 million+ tweets containing "iPhone" Collected using Topsy Pro (http://topsy.thisisthebrigade.com)
H&M	Facebook	01–2009 to 10–2014	~15 million data points from the official H&M Facebook page Collected using the Social Data Analytics Tool (Hussain & Vatrapu, 2014)

H&M revenues for Facebook data. That is, if a tweet/like can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. We extend Asur & Huberman (2010) in three important ways: (a) addition of Facebook social data, (b) theoretically informed time lagging of the independent variable, social media actions, and (c) domain-specific seasonal weighting of the dependent variable, sales/revenues. Figures 20.1 and 20.2 present the predicted vs. actual charts for Apple iPhone sales and H&M revenues respectively.

With regard to our prediction models, we observed a 5–10% average error from our predictive models with the actual sales and revenue data over three-year period of 2012–2014. In the case of the iPhone sales prediction model, our average error of 5% is not that far from the industry benchmark predictions of Morgan Stanley and IDC. That said, there are several challenges and limitations to the predictive analytics processes and their outcomes. 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

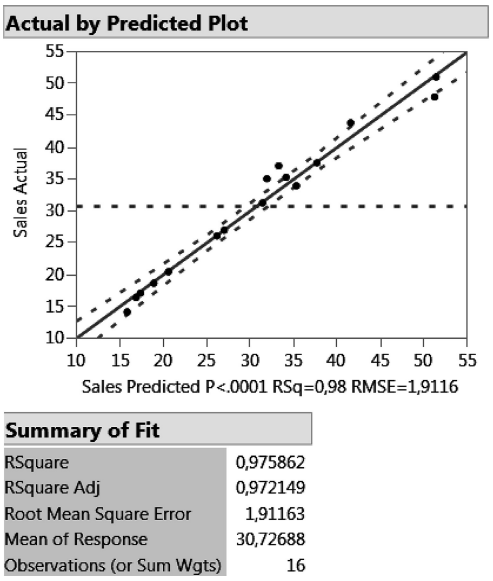

Figure 20.1 Predictive Model of iPhone Sales from Twitter Data

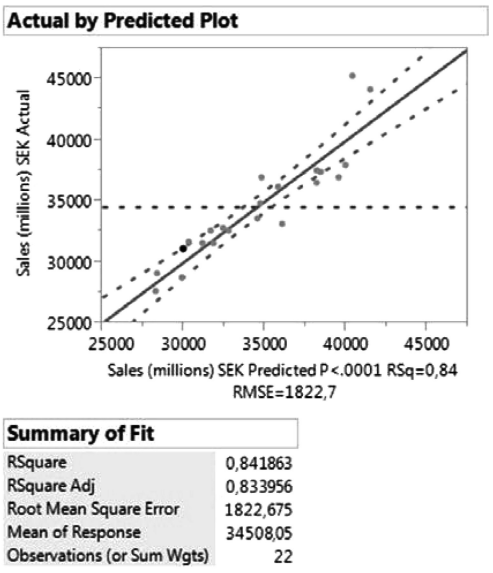

Figure 20.2 Predictive Model of H&M Revenues from Facebook Data

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.

CONCLUSION

Predictive models offer powerful tools as numerical forecasts and assessments of their uncertainty alongside quantitative statements more generally may improve decisions in companies and by public authorities.

The overall advice is to go for a parsimonious, simple model that captures the most important features of the data, that fulfils the model assumptions and that provides a good fit both in sample and out of sample. Furthermore, it is important that even during the phase where the model is applied for its purpose, it performance is still monitored. We present a general model for predictive analytics of business outcomes from social media data below.

$$y_t = \beta_a \times A_t + \beta_p \times P_t + \beta_d \times D_t + \beta_o \times O_t + \varepsilon_t$$

Where:

y_t = Outcome variable of interest

A_t = Accumulated time-lagged social media activity associated with outcome variable at time *t*

$$A_t = \Sigma A_{st}$$

A_{st} = Social media activity in terms of actions by actors on artifacts associated with outcome variable at time *t*

P_t = Individual or social psychological attribute(s) at time *t*

D_t = Social media dissemination factors

O_t = Other explanatory factors

A final word of caution will end this chapter: any predictive model is based on a certain set of information. It is necessarily backward-looking as it relies on historical data and irrespectively of how carefully the model specification and evaluation is done, there is no guarantee that the prediction of future values of the variable of interest will be reliable. The patterns or theories that the model relies on may break down and render the model useless for predictive purposes. That being said, careful predictive modelling is probably the best that can be done and, if applied and used following the state of the art with most emphasis placed on short term forecasting, predictive modelling is a very valuable tool.

ACKNOWLEDGEMENTS

We thank the members of the Centre for Business Data Analytics (http://bda.cbs.dk) for their feedback.

The authors were partially supported by the project Big Social Data Analytics: Branding Algorithms, Predictive Models, and Dashboards funded by Industriens Fond (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this chapter are those of its authors and do not represent the views of the Industriens Fond (The Danish Industry Foundation).

REFERENCES

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Belch, G. E., Belch, M. A., Kerr, G. F., & Powell, I. (2008). *Advertising and promotion: An integrated marketing communications perspective*: McGraw-Hill, London.

Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 91–94.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM, 11*, 450–453.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bothos, E., Apostolou, D., & Mentzas, G. (2010). Using Social Media to Predict Future Events with Agent-Based Markets. *IEEE Intelligent Systems*, *25*(6), 50–58.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(1), 1–20.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45. doi: 10.4269/ajtmh.2012.11–0597

Chung, J. E., & Mustafaraj, E. (2011). *Can collective sentiment expressed on Twitter predict political elections?* Paper presented at the AAAI.

Cioffi-Revilla, C. (2013). *Introduction to Computational Social Science: Principles and Applications*: Springer Science & Business Media.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). *Predicting the political alignment of Twitter users.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SocialCom).

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J.P., Sanchez, A., Nowak, A & Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal*, *214*(1), 325–346.

Council, N. (2013). Frontiers in massive data analysis: The National Academies Press Washington, DC.

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages.* Paper presented at the Proceedings of the first workshop on social media analytics.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). *Predicting postpartum changes in emotion and behavior via social media.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). *Characterizing and predicting postpartum depression from shared Facebook data.* Paper presented at the Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting Depression via Social Media.* Paper presented at the ICWSM.

Dijkman, R., Ipeirotis, P., Aertsen, F., & van Helden, R. (2015). Using Twitter to predict sales: a case study. *arXiv preprint arXiv:1503.04599*.

Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap, M. and Weeg, C. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159–169.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet Research*, *13*(4), e123.

Evangelos K, Efthimios T and Konstantinos T. (2013) Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559.

Gilbert, E., & Karahalios, K. (2009). *Predicting tie strength with social media.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). *The predictive power of online chatter.* Paper presented at the Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery in data mining.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social

media usage. *Computers in Human Behavior*, *28*(2), 561–569.

Hussain, A., & Vatrapu, R. (2014). Social Data Analytics Tool (SODATO). In M. Tremblay, D. VanderMeer, M. Rothenberger, A. Gupta, & V. Yoon (Eds.), *Advancing the Impact of Design Science: Moving from Theory to Practice* (Vol. 8463, pp. 368–372): Springer International Publishing, Switzerland.

Hussain, A., Vatrapu, R., Hardt, D., & Jaffari, Z. (2014). Social Data Analytics Tool: A Demonstrative Case Study of Methodology and Software. In M. Cantijoch, R. Gibson, & S. Ward (Eds.), *Analyzing Social Media Data and Web Networks* (pp. 99–118): Palgrave Macmillan, UK..

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts: https://www.otexts.org/fpp/

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.

Karabulut, Y. (2013). *Can Facebook predict stock market activity?* Paper presented at the AFA 2013 San Diego Meetings Paper.

Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008). *Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis.* Paper presented at the ECIS.

Lassen, N., Madsen, R., & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. *Proceedings of IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014), Ulm, Germany*, 81–90, ISBN: 1541–7719/1514, doi: 1510.1109/EDOC.2014.1520.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. and Jebara, T. Computational Social Science. *Science*, *323*(5915), 721–723. doi: 10.1126/science.1167742

Li, J., & Cardie, C. (2013). Early stage influenza detection from Twitter. *arXiv preprint arXiv:1309.7340*.

Mao, H., Counts, S., & Bollen, J. (2014). *Quantifying the effects of online bullishness on international financial markets.* Paper presented at the ECB Workshop on Using Big Data for Forecasting and Statistics, Frankfurt, Germany.

Radosavljevic, V., Grbovic, M., Djuric, N., & Bhamidipati, N. (2014). *Large-scale World Cup 2014 outcome prediction based on Tumblr posts.* Paper presented at the KDD Workshop on Large-Scale Sports Analytics , New York.

Ritterman, J., Osborne, M., & Klein, E. (2009). *Using prediction markets and Twitter to predict a swine flu pandemic.* Paper presented at the 1st international workshop on mining social media, Sevilla, Spain.

Robertson, S., & Vatrapu, R. (2010). Digital Government. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 44, pp. 317–364).

Robertson, S., Vatrapu, R., & Medina, R. (2010a). Off the wall political discourse: Facebook use in the 2008 US Presidential election. *Information Polity*, *15*(1), 11–31.

Robertson, S., Vatrapu, R., & Medina, R. (2010b). Online Video "Friends" Social Networking: Overlapping Online Public Spheres in the 2008 U.S. Presidential Election. *Journal of Information Technology & Politics, 7*(2–3), 182–201. doi:10.1080/19331681003753420

Sang, E. T. K., & Bos, J. (2012). *Predicting the 2011 Dutch senate election results with Twitter.* Paper presented at the Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, France.

Seiffertt, J., & Wunsch, D. (2008). Intelligence in Markets: Asset Pricing, Mechanism Design, and Natural Computation [Technology Review]. *Computational Intelligence Magazine, IEEE*, *3*(4), 27–30.

Shen, W., Wang, J., Luo, P., & Wang, M. (2013). *Linking named entities in tweets with knowledge base via user interest modeling.* Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3) 289–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., & Jiang, J. (2012). *Tweets and votes: A study of the 2011 Singapore general election.* Paper presented at the System Science

(HICSS), 45th Hawaii International Conference on System Sciences, Hawaii.

Tang, L., & Liu, H. (2010). Toward predicting collective behavior via social dimension extraction. *Intelligent Systems, IEEE*, *25*(4), 19–25.

Tsakalidis, A., Papadopoulos, S., Cristea, A. I., & Kompatsiaris, Y. (2015). Predicting elections for multiple countries using Twitter and polls. *Intelligent Systems, IEEE*, *30*(2), 10–17.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM, 10*, 178–185.

Vatrapu, R., Mukkamala, R., & Hussain, A. (2014). *A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings.* Paper presented at the Computational Social Science Workshop at the European Conference on Complex Systems 2014, Lucca.

Vatrapu, R., Robertson, S., & Dissanayake, W. (2008). Are Political Weblogs Public Spheres or Partisan Spheres? *International Reports on Socio-Informatics*, *5*(1), 7–26.

Vatrapu, R., Hussain, A., Lassen, N. B., Mukkamala, R., Flesch, B., & Madsen, R. (2015). Social set analysis: four demonstrative case studies. *Proceedings of the 2015 International Conference on Social Media & Society*. doi:10.1145/2789187.2789203

Voortman, M. (2015). Validity and reliability of web search based predictions for car sales.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565–578.

Weeks, B. E., & Holbert, R. L. (2013). Predicting dissemination of news content in social media a focus on reception, friending, and partisanship. *Journalism & Mass Communication Quarterly*, *90*(2), 212–232.

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., & Kim, D. K. (2013). Predicting national suicide numbers with social media data. *PloS one*, *8*(4), e61809.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919–926.