This is the authors' version published as:

Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, Peter Gloor, (2013) "The power of prediction with social media", Internet Research, Vol. 23 Iss: 5, pp. 528 – 543

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source.

http://dx.doi.org/10.1108/IntR-06-2013-0115

The Power of Prediction with Social Media

- Harald Schoen, University of Bamberg (Germany), harald.schoen@uni-bamberg.de
- Daniel Gayo-Avello, University of Oviedo (Spain), dani@uniovi.es
- Panagiotis Takis Metaxas, Wellesley College and Harvard University (USA), pmetaxas@seas.harvard.edu
- Eni Mustafaraj, Wellesley College (USA), eni.mustafaraj@wellesley.edu
- Markus Strohmaier, Graz University of Technology (Austria), markus.strohmaier@tugraz.at
- Peter Gloor, MIT (USA), pgloor@mit.edu

Abstract

Social media today provide an impressive amount of data about users and their societal interactions, thereby offering computer and social scientists, economists, and statisticians – among others– many new opportunities for research exploration. Arguably, one of the most interesting lines of work is that of predicting future events and developments based on social media data, as we have recently seen in the areas of politics, finance, entertainment, market demands, health, etc. In fact, an average of one in seven research papers presented at the WWW, ICWSM and IEEE SocialCom Conferences between 2007 and 2012 contain the term "predict" in their title. This upward trend, starting from 0 in 2006 and reaching 18% in 2012, shows a significant interest of the research community in predicting with Social Media.

But what can be successfully predicted and why? Since the first algorithms and techniques emerged rather recently, little is known about their overall potential, limitations and general applicability to different domains.

Better understanding the predictive power and limitations of social media is therefore of utmost importance, in order to be successful and avoid false expectations, misinformation or unintended consequences. Today, current methods and techniques are far from being well understood, and it is mostly unclear to what extent or under what conditions the different methods for prediction can be applied to social media. While there exists a respectable and growing amount of literature in this area, current work is fragmented, characterized by a lack of commonly accepted evaluation approaches. Yet, this research seems to have reached a sufficient level of interest and relevance to justify a dedicated section.

This special section aims to shape a frame of important questions to be addressed in this field, and fill the gaps in current research with presentations of early research on algorithms, techniques, methods and empirical studies aimed at the prediction of future or current events based on user-generated content in social media.

Introduction

Human beings are fascinated with what will happen in the future and, indeed, we even associate intelligence with an ability to predict future events (Hawkins 2004). In ancient times, several techniques were invented including inspecting bird flights, haruspicy, and astrology. Later, predictions were done mostly through experts who had developed their own intuitions and methods of prediction. Unfortunately, such expert knowledge is idiosyncratic and cannot be automatized or even duplicated. In more recent times, the research community has developed much more sophisticated techniques that aim to predict future outcomes using data-based models. Such model-based forecasts have proved to be quite successful in predicting a diversity of outcomes including economic, societal, and political outcomes (e.g., Campbell 2008; Clements and Hendry, 2011; Silver 2012). Despite their general success, even these models cannot predict the future perfectly, because real-world outcomes can change in ways that are not anticipated by data-based models.

The advent of social media provides researchers with a new and rich source of easily accessible data about individuals, society and, potentially, the world in general. In particular, data from social media captures online behavior of users who communicate or interact on a diversity of issues and topics. It is the intent of this special section to focus on novel methods of prediction that are based on data harvested from social media. In recent years, such data has shown to be very popular with scholars interested in developing predictive models. With varying success, an emerging community of researchers has utilized social media data for a wide variety of purposes, for example, to predict stock market movements (e.g., Bollen et al., 2010), to predict announcements of flu outbreaks (Lampos et.al., 2010), to forecast box-office revenues for movies (Asur and Huberman, 2010) and even to predict election outcomes (Gayo-Avello, 2012), to name a few. The models and areas of application are diverse and, moreover, predictions based on social media data have also attracted considerable attention from the public through traditional and online media. These media are projecting an impression of social media as a widely accepted and reliable source of data for predicting future outcomes.

However, reality is more complicated than that. There are many theoretical and methodological issues in predicting future outcomes using social media data that are far from being settled, and deeper studies and experiments are required to discover the true potential of social media as a reliable source of data. While prediction represents a problem in a wide variety of scholarly fields, social media-based forecasts today receive significant attention. We thus consider it appropriate to discuss and reflect on the promises of social media based forecasts as well as the perils and pitfalls it is plagued with, and strategies to address these problems.

This special section aims to serve as a platform for these and related matters. Its intended audience are computer scientists, social scientists, economists, statisticians, and other researchers interested in the application of multidisciplinary approaches to exploit user-generated contents to better understand (and predict) societal behaviors. This issue includes three works approaching

such topics from different points of view: the credibility of data appearing in social media; the detection of unexpected phenomena as deviations of the "pulse" of social media; and a conceptual framework to survey the current body of research.

This guest editorial is organized as follows: First, we discuss the different approaches to build forecasting models. Then, we analyze how such models could be adapted to the special circumstances of social media and the caveats that apply (e.g., the pervasive need of machine learning methods to ensure the quality of data). Additionally, we discuss one peculiar idiosyncrasy of social media-based forecasting: the fact that sometimes it is not forecasting but nowcasting; i.e. the variable of interest is estimated in real time using online trails as proxies. After that, we briefly survey the most representative research conducted in the topic up to date, and introduce the papers published in this special section. The editorial concludes with some final remarks.

Different Types of Forecasting Models

A fundamental question we need to address in order to tackle with previous issues is: What enables prediction based on social media data? A first requirement is that the prediction itself somehow must be encoded within the data; without any signal w.r.t. the phenomenon of interest the data would be rendered useless. Second, the data collection needs to maintain the encoding of the answer. Third, the analysis performed on the collected data is able to reveal the prediction. Without all three of these fundamental requirements, predictions are either not possible or no better than pure chance.

It is fundamental, therefore, to examine the ways the research community conducts both the process of collecting data and its analysis. We observe that there are three prevailing practices: Data could be collected through past logs of experiences, and **statistical** models are often employed to make sense of them. Data could also be collected on demand. A traditional and direct way to do that is by using polling, asking the public directly for their opinion or behavioral intentions, as is done with **survey** models. However, social media provides an additional and indirect[1] way to collect data on demand. Researchers can unobtrusively approach social media to observe the public's behavior and then derive their intention or opinion from the observed behavior (e.g., using machine-learning techniques, cf. Bishop, 2006). When the interest lies in the users' opinion about the outcome of an event, rather than their intention with regards to it, the method is somewhat comparable to **prediction markets** models.

[1] Maybe it would be more appropriate to denote such kind of collection as unobtrusive or non-reactive; see, for instance, Janetzko (2008) for a broad introduction to this approach to measurement in social sciences.

In the following subsections we will discuss these different types of forecasting models we have seen in various fields. Without loss of generality, we use electoral predictions as a running example for the discussion.

Prediction Market Models

The prediction market model attempts to capitalize on the so-called "wisdom of crowds" approach (Surowiecki 2004). A large number of people give their best guesses for an outcome variable. In this respect, this approach is based on subjective evidence. Then, the individual guesses are aggregated in some way and the aggregate guess, according to this line of reasoning, will closely approximate the real outcome. This approach is underlying a host of prediction markets (e.g., Arrow et al., 2008; Rhode and Strumpf, 2004; Forsythe et al., 1992). Participants deal with assets that are linked to the quantity of interest, i.e. the occurrence of an outcome or a parameter, such as a party's vote share. Market prices are thus interpreted as predictions of the occurrence probability or another parameter of interest.

Prediction markets have been shown to be quite successful in predicting several outcomes (cf. Wolfers and Zitzewitz, 2004). At the same time, it has been pointed out that successful market-based predictions require certain preconditions to be met, including a sound market architecture guaranteeing the heterogeneity of participants (Surowiecki 2004). Moreover, some critics object that electoral markets simply mirror information available from other sources, i.e. election polls, and add no new information (Erikson and Wlezien, 2008a).

Survey Models

We refer to the second kind of forecast models as "survey models" because their approach is typical of election surveys that are sometimes used to predict election outcomes. In this model, an appropriate random sample from the people who might affect future outcomes is required. Then, the people included in this random sample are questioned about the ways in which they intend to act (for example, vote in an election or purchase consumer goods). Then, the distribution of behavioral intentions is interpreted as a forecast of the future outcome.

This procedure assumes that the sample is not biased and respondents' future behavior does not differ systematically from their stated intention (e.g., Perry, 1979; Rattinger and Ohr, 1993). The usefulness of this model thus critically hinges upon the quality of the sample, the right questions to be asked in the survey, and the interval between the interview and the future outcome. Quite obviously, undecided respondents are a source of potential obstacles in the analysis.

Statistical Models

The third method builds on statistical models of the outcome of interest[2]. Using some kind of time-series analysis, univariate models aim at detecting past regularities in the outcome variable are then used to predict its future development. Multivariate models capture the relationship between the outcome variable and several predictor variables. Whereas data-driven models simply aim at detecting empirical relationships, theory-driven models identify predictors that can be linked to the outcome in theoretically meaningful ways. In this vein, the vote share of an incumbent party might be modeled as a function of the state of the economy several months before an election, the results of trial heat polls, and the length of incumbency (for a variety of electoral forecast models see, e.g., Campbell 2008; Erikson and Wlezien, 2008; Hibbs, 2008; Holbrook, 2008). Having established a robust empirical model, predicting a future outcome requires filling in relevant information on predictor variables and then calculating the dependent variable (for a discussion of statistical issues in predictions see, e.g., Brandt et al., 2011; Montgomery et al., 2012).

The success of statistical predictive models hinges upon the robustness of the empirical relationships, in particular, the patterns detected in the past are assumed to hold in the future. In the absence of a structural break, such predictions are likely to prove valuable. Yet, the absence of structural breaks cannot be taken for granted. For example, an external shock might alter the relationship between gross income and gross demand, or a new party might change the logic of party competition and vote choice. Put differently, the success of statistical models crucially hinges upon the assumption that the future closely resembles the past.

Forecasting Models with Social Media

In principle, the identified types of forecasting models can be adapted to and/or used in the context of social media. Yet, it remains unclear what type of model best suits the characteristics and the fabric of social media data.

Social media allow users to interact, to share content, and to create content collectively (O'Reilly, 2005; Shirky, 2008; Gauntlett et al., 2011). Social media comprise, inter alia, weblogs, social networking sites, and platforms for music, video, and photo sharing. Every move users make on social media is documented on machine readable formats. When analyzing these data, their origin must be taken into account. In particular, Internet users and, even more so, users of social media have voluntarily decided to use these applications and thus differ from the

[2] This approach, also labeled "econometric models", is common place, for instance, in economics (e.g., Clements and Hendry, 2011; Hendry and Ericsson, 2001; Granger et al., 2006) or electoral forecasting (e.g., Lewis-Beck and Rice, 1992; Campbell and Garand, 2000).

population at large in terms of demographic characteristics, socio-economic variables, and socio-political attitudes (e.g., Hargittai and Hinnant, 2008).

The characteristics of social media affect the three above-mentioned forecasting methods differently. For example, **prediction market** models fit nicely with certain characteristics of social media. Prediction markets can, quite straightforwardly, be conceived as a social media application. Social media connect a large set of people around the world, thereby increasing the number of potential participants in prediction markets. Moreover, social media might increase the diversity of participants, thereby potentially improving the quality of predictions. Since the success of prediction markets –at least in theory– depends on the market architecture, considerable attention should be paid on market design issues. When publishing their results, it is of utmost importance to report decisions concerning market design issues, including resistance to tampering, as they might influence prediction outcomes. Empirically, in the social media era many prediction markets on economic, societal, and political outcomes (e.g., Berlemann and Schmidt, 2001; Jacobsen 2000; O'Connor and Zhou, 2008; Pennock et al., 2001; Polgreen et al., 2007) as well as on sports events (e.g., Gil and Levitt, 2007) were established. More recently, traditional social media sites have implemented prediction markets (Qiu et al., 2011).

Compared to prediction markets, the **survey** model faces certain challenges when applied in the context of social media. While online surveys can certainly be employed on social media platforms, it remains unclear whether social media-based survey results are well-suited for predicting future outcomes. A valid survey-based prediction requires an unbiased sample from the relevant population and valid answers, i.e. behavioral intentions that closely resemble future behavior. An obvious obstacle to reliable predictions is the self-selection nature of social media users. So, even if one were to ask a large number of social media users for behavioral intentions, a prediction of an outcome in the population at large is likely to be biased. Yet, repetitive surveys might prove useful in predicting the direction of change of the variable rather than its absolute value. From a different perspective, the importance of sampling and timing suggests that these decisions should be made carefully and scholars should scrupulously report them.

When it comes to using **statistical** models, there are no obvious obstacles to applying them to social media. Studying data might result in the detection of a statistical relationship between a social media-based measure (e.g., the number of likes on Facebook or sentiment analysis scores) and the outcome of interest (e.g., economic growth or presidential approval rates, as in O'Connor et al., 2010). Having established a model, one just needs to fill in appropriate information on the predictor variables to forecast the outcome. The success of this endeavor however depends on the robustness of empirical patterns. Given a reasonable theoretical account of the relationship between predictor variables and outcomes, the odds of predictive success can be quite high. Understanding underlying mechanisms permits scholars to identify conditions of predictive success and distinguish substantive and (presumably) stable relationships from spurious ones. For example, predicting the output of a factory from its input is a rather safe bet. The more tenuous the theoretical link from predictors to outcome, the more unstable the empirical

relationship is likely to. To give an example: while the number of Twitter followers might be linked empirically to the number of votes a candidate receives, the causal link from predictor to outcome is rather weak. Accordingly, it is wise to check the robustness of a model repeatedly over long periods before using it for predictions. Nevertheless, even extensive calibrating and testing cannot guarantee predictive success if structural breaks occur.

The logic underlying statistical models provides scholars with a considerable leeway in establishing prediction models. Social media data comprise a host of information that might serve as predictors. They might be tweets, Facebook posts, or contents of weblogs. Assuming we have decided to utilize tweets, a host of additional questions has to be addressed. These include questions concerning the period and method of data collection, the preparation of raw Twitter data for prediction, the procedure to predict the outcome of interest, and the calibration and testing of the model that is to be used for out-of-sample predictions. Each of these questions provides scholars with a considerable leeway and each decision can critically affect the quality of the model. As a consequence, scholars should make these decisions deliberately and they should spend considerable effort to carefully report the details of choosing observations and time ranges, selecting relevant variables, and testing the model (e.g., Gayo-Avello, 2012; Jungherr et al., 2012). One way of achieving that would be making materials publicly available (e.g., King, 1995).

Additional Caveats regarding Social Media-Based Forecasting

Handling Noise and Bias in Social Media Data

Social media provides a fluid, instantaneous, cheap, unstructured way of collecting data at large scale. This has allowed researchers from very different backgrounds to exploit it for predictive purposes. Because of such diversity, all of the aforementioned predictive models have been applied with variable success. However, despite of the model of choice, social media poses problems regarding the quality and credibility of the collected data, and those problems must be addressed with techniques which are independent of the predictive models.

Using again an electoral example, one could approach the problem of predicting the outcome of a given election in, at least, the following ways:

1. Data from users expressing their opinion about the various parties and candidates is collected; these opinions are interpreted as vote intentions and, on the basis of that interpretation, a prediction about the actual results is made. This model would be an adaptation of the **survey** model to social media. In fact, this approach has appeared often in the literature of electoral prediction from social media.
2. A different approach could consist of collecting data from users expressing their opinion, not about parties or candidates but about the probable outcome of the elections. In other words, one could exploit social media data as a kind of predictive **market**.

3. Yet another way would be to correlate time series (obtained, e.g., from pre-electoral polls) with the evolution of different indicators collected from social media; this way, a model would be obtained that could estimate a poll conducted on election day. This approach would resemble a **statistical** model.

Unfortunately, all of those approaches are making assumptions about both social media users and the contents they produce that are difficult to hold across elections. For instance, the survey approach assumes that social media users are a representative sample of the population, there is no self-selection bias (e.g., sympathizers of each party produce contents more or less at the same rate), every user is a potential voter, and every piece of social content can be accurately interpreted and matched with a voting intention. The prediction market approach makes similar assumptions, particularly that users are expressing their actual opinions instead of trying to "cheat" the market. This sensitiveness to the credibility of data is also shared by the rest of models, including the statistical ones.

The fact is that, at least today, social media users are not a representative sample of the population; there is significant self-selection bias and, indeed, most of the contents are produced by very vocal minorities (e.g., Mustafaraj et al., 2010); automatically interpreting the opinion expressed in social media contents is far from easy; and, of course, some users are trying to "cheat" by either spreading misinformation (e.g., Metaxas and Mustafaraj, 2010 and 2012) or by producing an abnormally high volume of conversation by means of automated accounts (e.g., Ratkiewicz et al., 2011).

The sources of bias in social media are difficult but not necessarily impossible to handle. Demographic bias, for example, could be reduced by weighting contents accordingly to the strata to which each user belongs by using user profiling. Determining the trustworthiness of contents collected from social media, or the automated nature of some accounts would be unavoidable. Finally, reliable sentiment analysis should be applied in those cases where opinion is considered a key factor within the forecasting model.

Machine learning methods are commonly used to face such tasks and, hence, any researcher or practitioner interested in the area of social media-based forecasting should be familiar with them (cf., Bishop, 2006). Examples of its application to the aforementioned tasks can be seen in Pennacchiotti and Popescu (2011), Castillo et al. **(???? this issue)**, or Ratkiewicz et al. (2011). Finally, the reader interested in state-of-the-art sentiment analysis should consult the works by Pang and Lee (2008) or Liu (2012).

Aggregate vs Individual Predictions

All of the aforementioned predictive models and all of the research surveyed in this editorial aim to make predictions at the macro level. That is, the goal is to determine an outcome by aggregating data from huge numbers of users' recorded behaviors (e.g., making decisions or sucumbing to illness). However, predictions from social media can also be made at the micro

level; in other words, models can be developed to try to determine the underlying intention, the future behavior, or latent attributes of individual users on the basis of the trails they left in online media.

It must be noted that such kind of individual predictions are well-suited for machine learning approaches but are outside of the scope of this editorial. The interested reader should consult a number of recent works in this regard, such as Golbeck et al. (2011) and Quercia et al. (2011) on the prediction of personality traits; Choudhury et al. (2013) on the prediction of depression risk; Song et al. (2010) on the predictability of mobility patterns of individuals.

Nowcasting

An unstated assumption in the discussion so far was that the goal were to predict a future event or to forecast the evolution of a given variable in the future; the further in time, the more valuable the prediction. However, this does not have to be the only application of social media data. In fact, a number of researchers have been working not on forecasting future events but on so-called "nowcasting" or "predicting the present". That is, the current magnitude of a given variable that cannot be directly measured in real time is estimated indirectly. A number of works have been conducted on the feasibility of using social media and other user-generated trails as a source of data for nowcasting algorithms (e.g., Choi and Varian, 2009; Lampos and Cristianini, 2010 and 2012; or Signorini et al., 2011).

The main difference between nowcasting and forecasting models is that the later provide some "lead" time in the prediction, while the former provide a current estimate of a variable which cannot be measured in real time. Therefore, the only relevant question to decide between nowcasting and forecasting methods is whether lead time is really needed or if, in contrast, obtaining real time estimates can be valuable enough for the scenario at hand. Because of this, no distinction is made in the following literature survey between forecasting and nowcasting studies; however, we will clearly state whether the method is predicting the future or estimating the present.

A Brief Literature Review on Social Media-Based Forecasting

While we are not intent to provide an extensive survey of the literature in this editorial, we point to the paper by Kalampokis et al. in this issue **(???)** for an excellent work in that regard. Nevertheless, a brief –and therefore incomplete– review is needed to provide a broad picture of the state of the art. Below, we will examine studies of a few representative areas of prediction, namely that of influenza incidence, product sales, stock market movement, and electoral results. The order of presentation is by decreasing predictive success. This order is, arguably, closely related to desire by social groups of influencing the outcome, in ways that models have difficulty to anticipate. We acknowledge that other areas have been subject of social media-based forecasting, though with a much lower degree of attention from the research community.

Influenza Incidence

Forecasting the incidence of flu –both seasonal and H1N1– has been a recurring goal of researchers since 2008. That year, two widely cited papers were published showing that the evolution of queries submitted to major search engines exhibited a significant predictive power regarding the evolution of influenza cases and deaths across the U.S.[3] Polgreen et al. (2008) worked on influenza related queries submitted to Yahoo! and produced models to predict (from one to ten weeks before the official CDC announcement) the percentage of both influenza cultures and deaths due to influenza and pneumonia. And Ginsberg et al. (2009) worked on data provided by Google to develop a fairly similar model to predict the percentage of physician visits due to influenza-like illnesses. Their model was able to predict such cases one week in advance.

The Ginsberg et al. (2009) work was the basis for the Google Flu Trends website[4], a tool to show in real time both the incidence of influenza in the U.S. (and other countries) and the prediction made using their model. This tool became pretty popular and sparked subsequent research on predicting influenza incidence and the evolution of the H1N1 pandemic using social media data. As representative of this line of research and commonly applied methods we will just refer to the work by Lampos and Cristianini (2010) and Signorini et al. (2011). Both studies are examples of nowcasting and rely on the contents and the location of tweets in order to build models to measure influenza incidence in each region –in the first case regions of the U.K. and in the second one in the U.S. The methods applied are quite similar to those developed by Polgreen et al. (2008) and Ginsberg et al. (2009), strengthening the argument that statistical models are suitable to predict or monitor the incidence of health issues, provided that extensive and reliable historical ground truth data is available for training. Given this success, we expect that social media-based forecasting of public health scenarios using statistical models will be a promising area of research in the future.

Product Sales

Another area with substantial research is that of predicting product sales, such as books, video games, and movie tickets. Most of the published approaches to date also follow the statistical modelling approach. The main challenges of the proposed methods lie in finding an automatic way to determine the best keywords and their associated weights to fit user-generated data to the ground truth available for training.

Seminal work in this area was conducted by Gruhl et al. (2005) who have shown that the evolution of blog posts over time exhibited positive correlation with book sales and, in some

[3] It must be noted that, to the best of our knowledge, it was Eysenbach (2006) who first demonstrated the feasibility of using social media data to predict outbreaks of influenza, by devising a smart application of Google Adsense.

[4] http://www.google.org/flutrends/

cases, were able to predict (with days or even weeks of anticipation) spikes in sales. Similar work was conducted by Mishne and Glance (2006) correlating weblogs with box office performance during the opening weekend of a given movie by employing positive sentiment polarity towards each movie. This is of interest because sentiment analysis has become inextricably associated with social media-based prediction –although up to now it has been applied under the form of very simple methods.

Asur and Huberman (2010) were the first to use Twitter data for predicting film box office revenues during the opening and second weeks of each movie, and they found that tweet-rate time series exhibited strong correlation with movie earnings. When tweet-rates were combined with information about the number of theaters scheduled to release the film, the prediction exceeded that of the well-known Hollywood Stock Exchange prediction market, the gold standard of the industry. They also analysed the impact of sentiment analysis in the prediction, finding that it has some impact, especially after the movie has been released, but not as strong as that of tweet rates. Later, Wong et al. (2012) raised some doubts on the feasibility of predicting movie performance by mining Twitter data. Unfortunately, their methods are quite different from those applied by Asur and Huberman and, thus, in the absence of further research, it is difficult to ascertain if and under what conditions consumer patterns can be predicted from social media.

As in other cases, Web search query logs were exploited as social media data to forecast consumer behavior, namely retail, automotive and home sales, plus touristic visits (Choi and Varian, 2009). Goel et. al. (2010) included publically available Web search search data to dramatically increase the performance of baseline statistical models forecasting film box-office revenues, video game sales and the rank of songs on the Billboard Hot 100 chart.

Stock Market Movement

The stock market is, in some sense, one of the more promising areas to apply statistical modelling and, indeed, there is a rather large body of work regarding its predictability using social media data. Probably reflecting its financial importance, research on predicting the stock market from Web data largely predates the existence of user-generated content. An early work by Wüthrich et al. (1998) exploited contents from online financial newspapers to predict the closing values of stock markets in Asia, Europe and America. They find that a simple trading strategy informed by their prediction method –even when making a number of wrong predictions– would obtain a larger capital appreciation than that by the analyzed stock markets.

Tumarking and Whitelaw (2001) conducted a similar research but, instead of exploiting online financial news, they exploited a form of social media, in particular, a message forum specialized in financial topics. Their conclusions are somewhat contradictory to those of Wüthrich et al. (1998) as for the causality of the correlation. According to them, messages were not predicting or influencing the market and, *"if anything, the causality appears to run from the market to the financial forums."* In a similar vein, Antweiler and Frank (2004) have shown that message

boards do not predict stock returns although they help to predict volatility. In other words, while forum content is not mere noise, the available data is not feasible for developing simple predictive models.

More recent research provides a more optimistic outlook on the matter and tends to work with content generated by unspecialized users (as opposed to content produced in financial forums or websites). For instance, Choudhury et al. (2008) describe promising results when exploiting content from weblogs to predict stock market movements. Other researchers (e.g., Gilbert and Karahalios, 2010; Bollen et al., 2011a and 2011b; or Zhang et al., 2011) have shown that public mood measured through social media (e.g., LiveJournal or Twitter) is somewhat related to stock market movements: for instance, anxiety or fear tend to predate downward movements in the markets. It must be noted, however, that it has not been demonstrated that such emotional indices can predict stock returns.

Given the potential benefits of any successful model to predict stock markets, the feasibility of applying statistical models, and the availability of ground truth data, this is an area that will see much more research in the future.

Electoral Results

Predicting election outcomes from social media, in particular using Twitter data, has become quite popular. A considerable number of scholars have used tweets as indicators of electoral outcomes in different countries. Often, these studies utilize the number of mentions of a party or a candidate on Twitter before an upcoming election as an indicator of the vote share that party or candidate will receive (see for an overview Gayo-Avello, 2012).

Although some authors relate their methods to the wisdom of crowds, those predictions do not really belong to the prediction-market type. Prediction markets require participants to give best guesses of the likely outcome, not to just mention parties or candidates for any reason. Instead, their methods can be considered as resembling more closely the survey model. In election polls, respondents are asked to state the party or candidate they are likely to vote for. Then, mentions are counted and the proportion of mentions is interpreted as an estimate of the vote share a party or candidate will receive. Likewise, in social media predictions, party and candidate mentions are counted and the proportion of mentions is interpreted as an estimate of the vote share a party or candidate will receive.

At first glance, both procedures resemble each other quite closely. A closer look, however, reveals significant differences. For one thing, social media users are not asked to state their voting intention. Rather, they mention a party or candidate for any reason. When mentioning a party, they might praise it, criticize it or be neutral towards it. So, social media data do not necessarily reflect voting intentions. In addition, social media users certainly do not represent a random sample from the electorate, while professional survey analysts make considerable efforts

to draw a random sample from the electorate. In effect, social media data can differ dramatically from election survey data.

Another issue with most of those approaches is that they do not hypothesize and test empirical relationships between social media and election results[5] to use them in predicting the outcome of future elections. Fortunately, there are a few recent examples of more systematic approaches and, rather unsurprisingly, such works have relied on statistical models. For instance, Shi et al. (2012) and Lampos et al. (2012) describe similar methods to train regression models over pre-electoral polls conducted during the early stages of the campaign; then, such models can produce predictions from the evolution of different trends in Twitter. Their results are promising but were only tested on pre-electoral polling data, not the actual election results; moreover, their models were developed for one single election and, hence, it is unknown if they could be applied to future elections. Franch (2013) follows a similar approach combining data collected from many different social media sites, not just Twitter. Recently, Granka (2013) proposed an electoral forecasting model for the US based on web search data and past electoral results, not a single election. Finally, electoral predictions based on Web search volume data are also unsuccessful so far (Lui et al., 2011).

Therefore, electoral forecasting from social media is a field where further research is needed, and where the main focus should be put on providing general models that could be applied not to a single election but to multiple elections.

About the Articles Published in this Special Section

This special section includes the works by Castillo, Mendoza and Poblete (2013); Kalampokis, Tambouris and Tarabanis (2013); and Jungherr and Jürgens (2013). All of them approach the topic of predictions based on social media data from different –yet complementary– points of view.

Castillo et al. tackle the problem of credibility assessment of user-generated contents, in particular during emergency situations. As aforementioned, automatically determining whether a message is trustworthy or not is crucial when adapting any predictive model for social media-based forecasting. Their approach consists of a cascade of supervised classifiers that first determine whether a message is relevant to the topic of interest and then labels the message as credible or not. As shown in that paper, the performance of their method is pretty good and it should be easily adaptable to other scenarios.

Kalampokis et al. provide a thorough survey of the body of work regarding social media-based forecasting. To that end, they propose a conceptual framework within which current research is

[5] Except for the vague assumption that the larger the number of tweets, blog posts, Facebook likes or friends a candidate has, the larger his or her voting rate.

decomposed and described. According to that framework, every approach to make predictions from social data consists of two phases: data conditioning and predictive analysis. During the first phase, data is collected, filtered, and values for the predictor variables are computed. During the second phase, the predictive method is chosen, a model is built and, finally, its performance is evaluated. In addition to this framework, the authors provide a faceted approach to the existing literature by means of the areas of application, the kind of social media data exploited, the kind of sentiment analysis applied (if any), or the evaluation approach. In this regard, we note a striking finding of this paper: there is a considerable number of studies which rely only on post facto explanations to support the purported validity of the predictive method. Clearly, future research should avoid this kind of dubious criteria for evaluation.

Finally, Jungherr and Jürgens describe a novel method to analyze time series obtained from social media in order to find "abnormal" intervals. Those intervals are detected when the collected data exhibits a large deviation from the forecasted normal state of the series. In addition to that, the authors show how such deviations tend to correlate with offline phenomena. In this regard, this paper provides new tools to those researchers interested in applying the statistical modelling approach for monitoring purposes.

Conclusion

Our editorial highlights the multitude of issues one faces when trying to make predictions from social media, and points out the many pitfalls. We first introduce a taxonomy of models that have been used in the past to predict future events using social media data: The prediction markets, survey, and statistical models. For each of these models we discuss their relative advantages and the particular problems that they have been applied in. Further, we describe four areas of predictions that have attracted research interest in the past with variable success: influenza incidence, product sales, stock market movement, and electoral results. Finally, we introduce the accepted papers in this special issue and describe their important contributions.

The taxonomy and the accepted papers represent simply a first step towards a more systematic exploration of the potential and limitations of social media-based forecasts. To better understand them, comparative analyses are needed as they are likely to permit the study of the conditions controlling predictive success and their underlying mechanisms. As in many venues of human interaction, predictive success with social media is unlikely to have a simple "black and white" answer, but a complex one that depends on a multitude of factors.

If there is anything that the human experience has taught us is that predicting the future is both highly desirable and extremely difficult. Is there a model that seems to be better suited for predicting using social media? If so, is it among the three models we present here?

While we do not have reasons to doubt it, we are consciously cautious about the validity of our own arguments regarding the future of forecasting using social media data. In the future, one

might identify the existence of a new, fourth prediction model that is made possible by the idiosyncrasies of social media. So far, however, given its pervasiveness across multiple domains and the many successes claimed by researchers, it seems that statistical models is the most fruitful approach.

Acknowledgements

The work of P. Metaxas and E. Mustafaraj was supported by NSF grant CNS-117693.

References

1. Antweiler, W., and Frank, M.Z. 2004, "Is all that talk just noise? The information content of internet stock message boards", The Journal of Finance, vol. LIX, no. 3, p. 1259-1294.
2. Arrow, K. J. et al., 2008: The promise of prediction markets. Science 320, 877–878.
3. Asur, S. and Huberman, B. (2010), "Predicting the future with social media", in Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).
4. Bishop, C.M. 2006, Pattern recognition and machine learning. Springer, New York.
5. Bollen, J., Mao, H., and Zeng, X.-J. (2011), Twitter mood predicts the stock market, Journal of Computational Science, 2(1), pp. 1-8.
6. Bollen, J., Pepe, A., and Mao, H. 2011, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
7. Brandt, P. T., Freeman, J. R., and Schrodt, P. A. (2011), Real Time, Time Series Forecasting of Inter- and Intra-State Political Conflict, Conflict Management and Peace Science, 28(1), pp. 58-80.
8. Campbell, J. E. (2008). The trial-heat forecast of the 2008 presidential vote: Performance and value considerations in an open-seat election. PS: Political Science and Politics 41(4), 697–701.
9. Campbell, J.E., and Garand, J. (Eds.) 2000, Before the Vote: Forecasting American National Elections, Sage Publications, Inc. Thousand Oaks, California, US.
10. Castillo, C., Mendoza, M., and Poblete, B. (???? this issue), "Predicting Information Credibility in Time-Sensitive Social Media", Internet Research, vol. ??, no. ??, pp. ??
11. Choi, H., and Varian, H. 2009, "Predicting the present with google trends", Google TR, available at: http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf
12. Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. 2013, "Predicting Depression via Social Media", in Proceedings of the 7th International AAAI Conference on Weblogs and Social Media.
13. Choudhury, M., Sundaram, H., John, A., and Seligmann, D. 2008, "Can blog communication dynamics be correlated with stock market activity?", in Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, pp. 55-60.

14. Clements, M. P. and Hendry, D. F. (Eds.) (2011), The Oxford handbook of economic forecasting, Oxford University Press, Oxford.
15. Erikson, R. S. and C. Wlezien, 2008a: Are political markets really superior to polls as election predictors? Public Opinion Quaterly 72, 190–215.
16. Erikson, R. S. and C. Wlezien (2008b). Leading economic indicators, the polls, and the presidential vote. PS: Political Science and Politics 41(4), 703–707.
17. Eysenbach, G. 2006 "Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance", in Proceedings of AMIA Symposium, pp. 244-248.
18. Franch, F. 2013 "(Wisdom of the Crowds)²: 2010 UK Election Prediction with Social Media", in Journal of Information Technology and Politics 10, 57-71.
19. Forsythe, R., Nelson, F.D., Neumann, G.R. and Wright, J. 1992: Anatomy of an experimental political stock market. American Economic Review 82, 1142–1163.
20. Gauntlett, David. 2011. Making is Connecting: The Social Meaning of Creativity, from DIY and knitting to YouTube and Web 2.0. Cambridge, UK u.a.: Polity.
21. Gayo-Avello, Daniel. 2012, A meta-analysis of state-of-the-art electoral prediction from Twitter data, arXiv:1206.5851 [cs.SI]
22. Gilbert, E., and Karahalios, K. 2010, "Widespread worry and the stock market", in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.
23. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. 2009, "Detecting influenza epidemics using search engine query data", Nature, vol. 457, no. 19.
24. Goel, S., Hofman, J., Lahaie, S., Pennock, D., and Watts, D. 2010, "Predicting Consumer Behavior with Web search", PNAS 107(41):17486--17490, October 12.
25. Golbeck, J., Robles, C., Edmonson, M., and Turner, K. 2011, "Predicting Personality from Twitter", in Proceedings of PASSAT/SocialCom, pp. 149-156.
26. Granger, C.W.J., G. Elliott and A. Timmermann (eds.) (2006), Handbook of Economic Forecasting, Amsterdam, North-Holland.
27. Granka, L. 2013, "Using Online Search Traffic to Predict US Presidential Elections", in PS: Political Science and Politics 46 (1), 271-279.
28. Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. 2005, "The predictive power of online chatter", Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 78-77.
29. Hargittai, E. and Hinnant, A. 2008, "Digital Inequality: Differences in Young Adults' Use of the Internet", Communication Research 35(5): 602-621.
30. Hawkins, J. (2004). On Intelligence. Times books. ISBN 0805074562.
31. Hendry, D.F. and N.R. Ericsson (eds.) (2001). Understanding Economic Forecasts. Cambridge, MA: MIT Press.
32. Hibbs, D. A. (2008), Implications of the 'bread and peace' model for the 2008 US presidential election, Public Choice, Vol. 137, pp. 1-10.

33. Holbrook, T. M. (2008). Incumbency, national conditions, and the 2008 presidential election. PS: Political Science and Politics 41(4), 709–712.
34. Jacobson, B., Potters, J., Schram, A., van Winden, F. and Wit, J., 2000: (In)accuracy of a European political stock market. The influence of common value structures. European Economic Review 44, 205–230.
35. Janetzko, D. 2008, "Nonreactive Data Collection on the Internet", in The SAGE Handbook of Online Research Methods.
36. Jungherr, A., Jürgens, P. and Schoen, H. (2012), Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., and Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment", Social Science Computer Review 30(2), pp. 229-234
37. Jungherr, A., and Jürgens, P. (???? this issue), "Forecasting the pulse: how deviations from regular patterns in online data can identify offline phenomena", Internet Research ?? (??), pp. ??-??
38. Kalampokis, E., Tambouris, E., and Tarabanis, K. (???? this issue), "Understanding the Predictive Power of Social Media", Internet Research ?? (??), pp. ??-??
39. King, G., "Replication, Replication", PS: Political Science & Politics 28, pp. 444-452.
40. Lampos, V., and Cristianini, N. 2010. Tracking the flu pandemic by monitoring the Social Web. In the Proceedings of the 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), pp. 411-416, IEEE Press.
41. Lampos, V., and Cristianini, N. 2012. "Nowcasting Events from the Social Web with Statistical Learning", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, article 72.
42. Lewis-Beck, M.S., and Rice, T.W. 1992, "Forecasting Elections", Congressional Quarterly Press, Washington D.C., US.
43. Liu, B. 2012, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.
44. Lui, C., Metaxas, P. and Mustafaraj, E. 2011, "On the Predictability of the US Elections through Search Volume Activity", Proceedings of the e-Society Conference, Avila, Spain.
45. Metaxas, P. and Mustafaraj, E. 2010, "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search", In Proceedings of the WebScience10: Extending the Frontiers of Society On-Line, April 26-27th, 2010, Raleigh, NC: US.
46. Metaxas, P. and Mustafaraj, E. 2012, "Social Media and Elections", Science, Policy Forum, October 26.
47. Mishne, G., and Glance, N. 2006, "Predicting movie sales from blogger sentiment", In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.
48. Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012), Improving Predictions Using Ensemble Bayesian Model Averaging, Political Analysis 20(3), pp. 271-291.

49. Mustafaraj, E., Finn, S., Whitlock, C., and Metaxas, P. 2011, “Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail”, In Proceedings of IEEE PASSAT/SocialCom, pp. 103-110.
50. O’Connor, P. and Feng Z., 2008: The tradesports NFL prediction market: An analysis of market efficiency, transaction costs, and bettor preferences. The Journal of Prediction Markets 2, 45–71.
51. O’Reilly, T. 2005. „What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.“ O’Reilly Blog, September 30, 2005. http://oreilly.com/web2/archive/what-is-web-20.html (January 16, 2012).
52. Pang, B, and Lee, L., 2008, “Opinion Mining and Sentiment Analysis”, Foundations and Trends in Information Retrieval, vol. 2, no. 1-2.
53. Pennacchiotti, M., and Popescu, A.M. 2011, “A Machine Learning Approach to Twitter User Classification”, in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
54. Perry, P. (1979), “Certain problems in election survey methodology”, Public Opinion Quarterly, Vol. 43, No. 3, pp. 312-325.
55. Polgreen, Philip M., Forrest D. Nelson and George R. Neumann, 2007: Use of prediction markets to forecast infectious disease activity. Healthcare Epidemiology 44, 272–279.
56. Polgreen, P.M., Chen, Y., Pennock, D.M. and Nelson, F.D. 2008, “Using internet searches for influenza surveillance”, Clinical Infectious Diseases, vol. 47, no. 11, pp. 1443-8.
57. Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. 2011, “Our twitter profiles, ourselves: Predicting personality with twitter”, in Proceedings of PASSAT/SocialCom.
58. Qiu, L., Rui, H. and Whinston, A. B., A Twitter-Based Prediction Market: Social Network Approach (December 1, 2011). Available at SSRN: http://ssrn.com/abstract=2047846 or http://dx.doi.org/10.2139/ssrn.2047846
59. Rattinger, H. and Ohr, D. (1989), „Wahlprognosen in einer Welt ohne Stichprobenfehler: Analytische Überlegungen und empirische Befunde“, in: Falter, J.W., Rattinger, H., and Troitzsch, K.-G. (Ed.), Wahlen und politische Einstellungen in der Bundesrepublik Deutschland: Neuere Entwicklungen der Forschung, Peter Lang, Frankfurt, pp. 282-331. [article title: “Predicting election outcomes in a world without sampling error. Analytical considerations and empirical evidence”; volume: “Elections and political attitudes in the Federal Republic of Germany: Recent developments in research”]
60. Rhode, P.W. and K. S. Strumpf (2004), Historical Presidential Betting Markets. Journal of Economic Perspectives 18, 127-142.
61. Shirky, Clay. 2008. Here Comes Everybody: The Power of Organizing Without Organizations. New York, NY u.a.: The Penguin Press.
62. Signorini, A., Segre, A.M., and Polgreen, P.M. 2011, “The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic”, PLoS ONE 6(5): e19467.

63. Silver, N. (2012), The Signal and the Noise. The Art and Science of Prediction, London: Allen Lane.
64. Song, C., Qu, Z., Blumm, N., and Barabási, A.L. 2010, “Limits of Predictability in Human Mobility”, Science, vol. 327, pp. 1018-1021.
65. Surowiecki, J. (2004), The Wisdom of Crowds, Doubleday, New York.
66. Tumarkin, R., and Whitelaw, R.F. 2001, “News or noise? Internet postings and stock prices”, Financial Analysts Journal, vol. 57, no. 3, p. 41.
67. Wolfers, J. and Zitzewitz, E. (2004), Prediction Markets. Journal of Economic Perspectives 18, 107-126.
68. Wong, F.M.F., Sen, S., and Chiang, M. 2012, “Why watching movie tweets won't tell the whole story?”, in Proceedings of WOSN’12.
69. Wüthrich, B., Permunetilleke, D., Leung, S., Cho, Zhang, J., and Lam, W. 1998, “Daily prediction of major stock indices from textual WWW data”, in Proceedings of KDD-98.
70. Zhang, X., Fuehres, H., and Gloor, P.A. 2011, “Predicting stock market indicators through twitter ‘I hope it is not as bad as I fear’”, Procedia - Social and Behavioral Sciences, vol. 26, pp. 55-62.