EJASA:DSS
Electronic Journal of Applied Statistical Analysis:
Decision Support Systems and Services Evaluation

EJASA:DSS – Electronic Journal of Applied Statistical Analysis
http://siba-ese.unisalento.it/index.php/ejasa_dss/index
e-ISSN: 2037-3627
DOI: 10.1285/i2037-3627v5n1p4

Big Data or Big Fail? The Good, the Bad and the Ugly and the missing role of Statistics

By Iacus

Published: 29 December 2014

Electronic Journal of Applied Statistical Analysis
Vol. 05, Issue 11, December 2014, 4-11
DOI: 10.1285/i2037-3627v5n1p4

Big Data or Big Fail? The Good, the Bad and the Ugly and the missing role of Statistics

Stefano Maria Iacus*[1]

[a] *University of Milan, Department of Economics, Management and Quantitative Methods, Via Conservatorio 7, 20122 Milan, Italy*

Published: 29 December 2014

The so called “Big Data” are data which we think as being “big” because of their volume, their amount per unit of time and because they are unstructured. The usual sources of big data are administrative repositories, transaction data or social media and social network feeds. Someone defines big data as those data which cannot be analyzed on a desktop machine or stored on one’s hard disk. These ways of defining big data completely miss the point of view of Statistics: they seem to be tailored more to advertising campaign of SaS or storage solution rather than to Science. Moreover, recent big fails, like e.g. the famous/infamous Google Flu Trend experiment, raised a series of popular news paper articles against the validity of information contained in these data and Statistics itself, even though none of these bad practices has been conducted by statisticians. While Information Technology and Computer Science are good at efficiently retrive and manage them, these data should be soon brought back into the field of Statistics to where data belong and this Special Issues of EJASA is one important step in this direction.

keywords: big data, social media, unstructured data.

1 Introduction

Big Data are those data where “Big” is often capitalized for unknown reasons and whose characteristics are described in terms of three main aspects (Daas et al., 2012; Couper, 2013):

*Corresponding author: stefano.iacus@unimi.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1. "volume": they exceed the capacity of traditional computing methods to store and process them;

2. "frequency": data come from streams or complex event processing, size per unit of time matters;

3. "unpredictability" : data come in the many different forms, they are raw, messy, unstructured, not ready for processing, not relational, ecc.

Big Data are sometimes called "organic data" (Groves, 2011) to express the idea that these data are generated, most of the times, automatically by systems, computers or probes or humans interacting through digital devices. In fact, the main types of these data can be summarized in terms of their source as follows:

- administrative data: stored in data bases generated by persons or organizations for regulatory or other government activities. In most cases these data are confidential, but in the last years, with the advent of the "Open Data" philosophy, more and more data are becoming available for public use although none of these data is usually collected or designed for scientific purposes;

- transaction data: they are generated through high frequency financial transactions on the stock markets, e-commerce transactions or loyalty card-type transactions at shops, phone records, web surfing, and so forth; these data, originally collected to increase the user experience are clearly used for secondary scopes like marketing and profiling in general. While these data may have a variety of different structures, the structure itself is well defined because they are collected for specific purposes. Their main issues consist in how to relate them;

- social media or social networking data: these data are created by explicit actions of people and by interactions with others. The purpose of these data is different and depends on the user and the social media, unpredictable in terms of time, space, volume and content, and highly unstructured (for the same user they can be: "likes", pokes, winks, photos, quotes, ecc), identity and unicity of the user is hard to assess. Nevertheless, these data convey a lot of information for the social scientists.

2 Big data and big mistakes

The Guardian in 2011[1] revealed that U.S. government manipulated social media by creating fake accounts with the aim of doing automatic pro-American propaganda. It is not so unfrequent to read about VIPs or politicians having fake followers, like Justin Bieber's 37 millions of followers[2]. By a small error, it is estimated that 83 millions of

[1] `http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks`

[2] `http://www.digitalspy.com/music/news/a471915/justin-bieber-twitter-followers-50-percent\ \-are-fake-says-report.html`

Facebook accounts (8.7% of the total accounts) are fake, half of them just duplicates of the same accounts. In March 2014, *The Financial Times*[3] published an article entitled *"Big data: are we making a big mistake?"* focused on the big fail of the Google Flu Trends experiment (Cook et al., 2011; Butler, 2013) and many other similar examples. These articles report the great disillusion of the naïve idea that the volume of data itself can replace the scientific reasoning (Anderson, 2008), like the enthusiasm (Noble, 2003) raised by the early completion of Human Genome Project in 2003, i.e. the mapping of all genes in the human DNA. Despite having the exact mapping of all genes, those data as is, have no information *per se*: once the human genome was transcripted biologists looked at it with no clue of its real meant. The same applies to the sea of social network discussions: we can download them all but they appear to be just noise. Having the data and being able to store, manipulated and move them does not coincide with the ability of extracting information from them.

3 Bringing back data to where they belong: Statistics

The previous "big fail" examples and many other come from researchers in fields outside of, although not unrelated to, Statistics. While computer scientist and information technology experts know very well how to optimize, store, query and distribute task over what we call big data, most of the time when they analyze them in a typical statistical framework, they base their work on the application of standard technologies like machine learning, data mining, ecc, to these data. This fact has so far produced the above mentioned bad practices so that even part of the statistical community use the same examples to complaint against the use of big data for serious research. Clearly, the problem is not in the data but in the analysts. In other words *"there are big data and small statisticians"*.

3.1 Change the statistician, not the data!

Without being too generic, let us focus on the third type of source of big data: the social media data. These are, by far, the most unstructured data, volatile (people appear and disappear), fast growing, high frequency and so forth. In addition to that, these data are mainly composed of digital texts expression of natural and informal language and contain most noise. In this field it is quite rare to have a benchmark to assess the validity of the analysis but at least in the case of the electoral competition. This is an extremely rare case in which it is possible to have a precise answer to the research question after the event occurred just by counting the votes.

Many authors have shown how social media data produce results close to the final outcome. For example, Lindsay (2008) developed a sentiment classifier based on lexical induction and reported correlations between polls and the content of wall posts available on Facebook during the 2008 US presidential election. Similarly, O'Connor et al. (2010) confirmed that Obamas approval rating is correlated with the sentiment expressed by

[3] http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html

Twitter users. Furthermore, sentiment analysis performed as well as polling in predicting the results of the 2011 Dutch senate election (Tjong Kim Sang and Bos, 2012), while analyses of multiple social media sites (Facebook, Twitter, Google, and YouTube) outperformed traditional surveys in estimating the results of the 2010 UK election (Franch, 2012).

Other successful attempts to forecast elections in several countries are available as well, e.g. Ireland (Bermingham and Smeaton, 2011), Portugal (Fonseca, 2011), Singapore (Choy et al., 2011; Skoric et al., 2012), the Netherlands (Sanders and van den Bosch, 2013), the United Kingdom (Tweetminster, 2010) and the United States (Choy et al., 2012; DiGrazia et al., 2013; Jensen and Anstead, 2013; Shi et al., 2012).

The main criticism (Gayo-Avello, 2012; Jugherr et al., 2011; Metaxas et al., 2011) of the above studies, is that they have been conducted after the election was held in most cases. When the forecast has been a real one, the criticisms was based on the fact that the result were not reproducible on the same (Twitter or Facebook) data by using the best practice from machine learning or natural language processing field. Ceron et al. (2014) performed a meta-analysis putting together all available results for more than 20 elections and came to the natural conclusion that what matters is the method of analysis and not solely the data source or their volume. The finding of this analysis were that the best practice in text analysis, and in particular, of those texts coming from the social media, need to satisfy some elementary statistical principles:

- "look into the data" or "use supervised statistical methods": data coming from social media contain lot of noise in the sense that, most of the times, the selected texts can be considered as *off topic* for the purpose of the analysis. A blind application of ontological dictionaries, the counting of mentions of a candidate (or a party), the count of followers, the number of likes or retweet, are totally biased measures of the willingness to vote for a candidate. Explicit intention to vote must be captured by hand-coding of a sample of texts and then this manually tagged set (or the *training set* as we usually call it) can be used in supervised statistical methods later;

- avoid standard data-ming techniques in the first place: although designed to discover patterns in the data, they are not designed to extract semantic structure of a text; at most, an average clustering technique may produce aggregation based on the frequency of the words or the alike which convey no information for the above mentioned reasons.

- "do not perform individual classification": Support Vector Machines, Neural Networks, Random Forests, Classification Trees, logistic or multinomial regressions, and so forth are all based on the the assumption that the *training* set and the *test* set (the set for which we want to predict the categorization) come from a well specified population. But this assumption is an unrealistic one. From what already stated, the highest semantic category in social media data is usually, by far, the *off topic* category but *off topic* texts are not as such because they use completely different words from the text of the training set. This means that, even the best classifier, will attribute with high probability the outcome of the classification to the category *off topic* and very rarely the true semantic category. As a result, the individual miss-classification will be very high and the final aggregation of the estimated proportions of the different categories will be highly biased. For this reason, in this particular context it is better to estimate directly the aggregated distribution of opinions by using new *ad hoc* method designed explicitly for this field (Hopkins and King, 2010; Ceron et al., 2013).

When one of the above pillars of sentiment analysis is missing, the building will collapse into ruins, i.e. the output of the analysis will be the estimation of the noise (or the bias) rather than the estimation of the signal.

Open issues remains like representativity of the statistical unit, demographic information of the social media users, and so forth, or how to integrate these results with official statistics and standard survey methods, but this should not stop the statistical investigation based on social media data. It is the task of the statistician to think about possible solutions.

4 Conclusions

The previous example of sentiment analysis over big data is just one of the many we have seen in the recent years. For example, at the early stage of genomic analysis, we have seen many dumb applications of standard statistical or machine learning techniques to, e.g., microarray data. Of course gene expressions do contain information, but standard machine learning or data-mining technique will not reveal any. Indeed, joint multiple-testing techniques and other *ad hoc* methods have been developed to assess, based on correct statistical grounds, the functional relationships between genes and pathologies.

To summarize, when statistics comes to data or, better, when the statistician approaches data analysis, the main focus should not be the data *per se* or the model alone, it is always the correct statistical model given the data at hands. The blind application of machine learning, data-ming techniques or other well known methods may severely affect the results of the analysis no matter how good these techniques are for the specific task they were designed for.

Statistics is not merely a set of tools, is mostly a way of looking at the world. Statistics is how one thinks about problems and how he or she tries to solve them. Big data are “today’s data” and clearly every data is a different beast which requires different methodologies. But these new data are more and more accessible and even if they require powerful backends and advanced statistical engines, these can be easily managed today from everyone’s laptop even using open source tools. Computer Science and Information Technology are a must as well as interactions with many disciplines, but the general principle of looking into the data and thinking at the problem in a statistical way seems to be too seldom neglected.

It is time for the data to go back to Statistics or, better, for the statisticians to start looking at the big data in an inclusive way. Many signals of this change of perspective can be already spotted and I think this special Issue of the *Electronic Journal of Applied*

Statistical Analysis devoted to the topic is one further important contribution in this direction.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *16:07*, February 13:http://www.wired.com/science/discoveries/magazine/16--07/pb_theory.

Bermingham, A. and Smeaton, A. (2011). On using twitter to monitor political sentiment and predict election results. In *paper presented at the Workshop on Sentiment Analysis where AI meets Psychology*, November 13, Chiang Mai, Thailand.

Butler, D. (2013). When google got flu wrong: Us outbreak foxes a leading web-based method for tracking seasonal flu. *Nature*, February 13:http://www.nature.com/news/when–google–got–flu–wrong–1.12413.

Ceron, A., Curini, L., and Iacus, S. (2013). *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete.* Springer, Milan.

Ceron, A., Curini, L., and Iacus, S. (2014). Using social media to forecast electoral results. a meta-analysis. *UNIMI - Research Papers in Economics, Business, and Statistics. Statistics and Mathematics*, 62:http://services.bepress.com/unimi/statistics/art62.

Choy, M., Cheong, M., Ma, N., and Koo, P. (2011). A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *ArXiv*, page http://arxiv.org/abs/1108.5520.

Choy, M., Cheong, M., Ma, N., and Koo, P. (2012). Us presidential election 2012 prediction using census corrected twitter model. *ArXiv*, page http://arxiv.org/abs/1211.0938.

Cook, S., Conrad, C., Fowlkes, A., and Mohebbi, M. (2011). Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLoS ONE*, 6(8):e23610.

Couper, M. (2013). Is the sky falling? new technology, changing media, and the future of surverys. *Survey Research Methods*, 7(3):145–156.

Daas, P., Roos, M., van de Ven, M., and Neroni, J. (2012). Twitter as a potential data source for statistics. *Den Haag/Heerlen, The Netherlands: Statistics Netherlands*, (Discussion paper 201221).

DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PlosONE*, 8(11):e79449.

Fonseca, A. (2011). Modeling political opinion dynamics through social media and multi-agent simulation. In *First Doctoral Workshop for Complexity Sciences*, idpcc.dcti.iscte.pt/docs/Papers_1st_Doctoral_Workshop_15-6-2011/AntonioFonseca.pdf.

Franch, F. (2012). (wisdom of the crowds): 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10:57–71.

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94.

Groves, R. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871.

Hopkins, D. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Jensen, M. J. and Anstead, N. (2013). Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process. *Policy & Internet*, 5(2):161–182.

Jugherr, A., Jürgens, P., and Schoen, H. (2011). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan a, sprenger to, sander pg and welpe im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30(2):229–234.

Lindsay, R. (2008). Predicting polls with lexicon. In *April 14*, http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon.

Metaxas, P., Mustafaraj, E., and Gayo-Avello, D. (2011). How (not) to predict elections. In *Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, pages 165–171, Los Alamitos, CA, USA.

Noble, I. (2003). Human genome finally complete. *BBC News*, April 14:http://news.bbc.co.uk/1/hi/sci/tech/2940601.stm.

O'Connor, B., Balasubramanyan, R., Routledge, B., and N.A., S. (2010). From tweets to polls: linking text sentiment to public opinion time series. In *Proceedings of the fourth international AAAI conference on weblogs and social media*, Washington, DC, 2326 May.

Sanders, E. and van den Bosch, A. (2013). Relating political party mentions on twitter with polls and election results. *unpublished manuscript*, pages `http://ceur--ws.org/Vol--986/paper_9.pdf`.

Shi, L., Agarwal, N., Agrawal, A., Spoelstra, G., and Spolestra, J. (2012). Predicting us primary elections with twitter. *unpublished manuscript*, page `http://snap.stanford.edu/social2012/papers/shi.pdf`.

Skoric, M., Poor, N., Achananuparp, P., Lim, E., and Jiang, J. (2012). Tweets and votes: A study of the 2011 singapore general election. In *Proceedings of 45th Hawaii International International Conference on Systems Science (HICSS-45 2012), IEEE Computer Society*, pages 2583–2591, Los Alamitos, CA, USA.

Tjong Kim Sang, E. and Bos, J. (2012). Predicting the 2011 dutch senate election results with twitter. In *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks*, Avignon, France.

Tweetminster (2010). Can word-of-mouth predict the general election result? a tweetminster experiment in predictive modeling. pages `http://www.scribd.com/doc/29154537/Tweetminster--Predicts`.