THE EFFICIENCY AND RELEVANCE OF DATASETS

BENJAMIN Ben (PhD Student)

DICENT-IDF laboratory, University of Paris-EST, Marne la Vallée

DAVID Amos

DICENT-IDF laboratory, Lorraine University

CONTENT

GENERAL INTRODUCTION

According to *D. Agrawal et al [1]*, data **heterogeneity**, **scale**, **timeliness**, **complexity** and **privacy** are known issues with Big Data which always **impedes progress** at all phases of the Big Data pipeline or any process that can create value from Big Data.

In addition to these challenges, while focusing on data collection phase, we will point out some challenges with certain data sources.

We will also emphasize on how these challenges can negatively affect **data integrity** which in turn affects **data quality.** This finally affects the correctness of the findings or insights obtained from analyzing the resulting datasets negatively.

…GENERAL INTRODUCTION

Data quality:

Is the measure of set of features or parameters describing their ability to satisfy user's expectation in a given area of interest (context).

It is closely connected with both the data form and the value of information carried by the data.

Data integrity:

It is the measure of assurance of the accuracy and consistency both physically and logically of data over its entire life cycle.

Is one of the most important feature or parameters of data quality.

SURVEY DATA COLLECTED FROM HUMAN RESPONDERS

- According to M. Dapkus and J. Pridotkienė [7], about **50%** of customers filling survey forms may actually provide false response because they want to boast about their behaviours or defensive about their selves.

- Sometimes, some surveys may be asking people what they will like to do in the future.

- S. Stephens-Davidowitz, puts it that people provide false response to friends, lovers, doctors, surveys and themselves.

…SURVEY DATA COLLECTED FROM HUMAN RESPONDERS

- Companies deliberately lure their visitors or customers to fill these surveys in a particular way, which favours these companies' views and opinions. According to Kimble who said that the use of Facebook's like button, which is taken as an indication of approval, can easily be manipulated by offers such as

"like our product and enter a draw to win a luxury holiday"

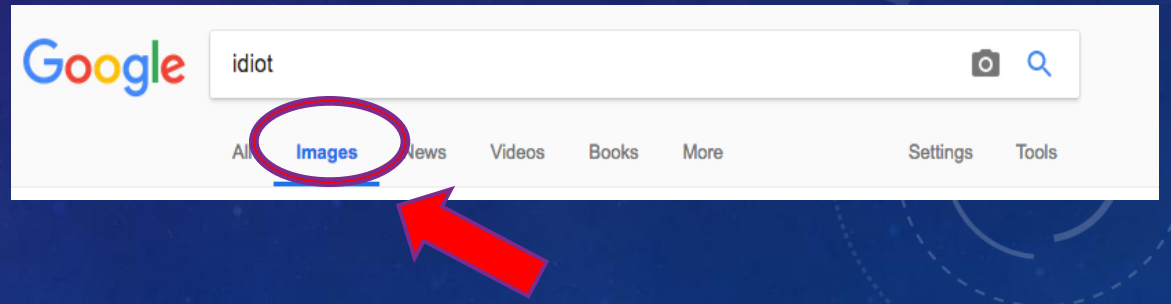
- People can genuinely forget the truth or make genuine mistakes while filling your survey forms.

SURVEY DATA COLLECTED BY AUTOMATED PROCESSES (SOFTWARE)

- Most websites especially online shopping sites, tend to acquire data from all their online visitors for the purpose of. Improving visitor's experience Targeted advertisement

Bill Stensrud, Chief Executive Officer of InstantEncore, "Everything I buy from Amazon is a present for somebody else, and so their recommendation engine is meaningless to me. Someday, they'll figure that out" **I don not always buy the same products twice!**

- Another example is the Google Bombing!



- Microsoft's TAY twitter bot

DETERMINING THE RIGHT DATA SOURCES AND THE MEANINGFUL VOLUME OF DATA REQUIRED

- “A data set may have many millions of pieces of data, but this does not mean it is random or representative”, Danah Boyd and K. Crawford [2]

  Big Data is now considered less about the data size but more about the capacity to search, aggregate, and cross-reference random and representative data sets

- Data scientist are not always there at the beginning when companies begin to automate their process and when the initial data modeling was done. This could lead to all sorts of problem ranging from missing data etc..

- Twitter, Facebook, etc. does not represent ‘all people’, and it is an error to assume ‘people’ and ‘Twitter or Facebook users’ are synonymous

VARYING INTEGRITY OF DATA SOURCES

- Apart from the physical and logical integrity of data, timing can affect data integrity negatively.

  Financial markets have huge data streaming in at real-time and several sources will present different varying figures for the same item due to delays in data acquisition and processing.

- According to Danah Boyd and K. Crawford [2], “Large data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together”.

SOCIAL MEDIA BIAS

- Humans are known to generally provide false information especially on social media platforms.

  Giving false information is part of human's day to day activities and this habit is what distinguishes us from other animals

- One of the aspects that is often overlooked when studying social media is the presence of bias in the data, F. Morstatter [5]

- Prince William (Duke of Cambridge), said social networks had allowed "misinformation and conspiracy to pollute the public sphere"

- **NSPCC** says the minimum age requirements for most social media platforms is 13. However, **Ofcom**, which is a regulator for the communications services in UK, found out that 46% of 11-year-olds, 51% of 12-year-olds and 28% of 10-year-olds now have a social media profile

DATA GENERATED FROM SENSORS

- There are two main primary sources of data, namely machine generated and human generated

- Cisco's Internet Business Solutions Group predicts that by 2020, 50 billion Integrated IoT (IIoT) devices will be deployed and active around the world.

- Wearable sensors have met both progress and setbacks. Some of these setbacks have led to unmet needs for doctors to continuously obtain medical quality data from their patients . J. Heikenfeld [6]

- Some of these challenges occur because these sensors are sometimes delicate devices and could be malfunctioning, wrongly calibrated, wrongly configured, wrongly installed or interference from the environment acting negatively on the sensor's readings or connectivity which leads to inaccurate sensor data.

CROWDSENSING AND CROWDSOURCING

- Crowdsensing is a paradigm used to describe a situation where people who have smart devices equipped with considerable number of sensors can participate in sharing the data generated by their smart devices.

- Crowdsourcing is the process of seeking and recruiting those with the required smart devices who are willing to participate in sharing the data captured from their smart devices

- This has major privacy and security issues which lead to poor participation.

"Even though we were told that there was a de-identification system in place, it didn't work. In theory, I think de-identification is an excellent approach, and it is one of the things that we EPIC continue to propose because it is one way to reconcile the public benefit while minimizing private harm. But it has to work." Marc Rotenberg

…CROWDSENSING AND CROWDSOURCING

According to Yang [4], “a user would not be interested in participating in crowdsensing, unless it receives a satisfying reward to compensate its resource consumption and potential privacy breach.”

“A problem that arises from the opt-in nature of crowdsensing applications is when malicious individuals contribute erroneous sensor data (e.g., falsified GPS readings); hence, maintaining the integrity of sensor data collected is an important problem.”, R. Ganti [9]

This issue was raised in 2011, but all the major reward mechanism said little or nothing about those who will take advantage of the financial rewards by upload falsified data in order to trick the system to pay them more and more financial rewards

PROBLEM IDENTIFICATION

- According to M. Rimando [8], “Data collection is the first stage in the research process.”.

- **Problem Identification** which is actually the process of defining the problem, establishing the problem domain and sometimes, the application domain also.

- Many times, data scientists are not involved in the problem definition stages and even in some situations, data scientists are not involved in building and selecting the data collection tools and data sources.

CONCLUSION

1. Problem Identification and problem domain definition should be considered as very important and the starting point

2. Survey data collected from humans may not be a 100% true reflection of responders views. This is also true for survey data collected by automated processes.

3. Right data sources and the volume of data required depends on the problem identified and the domain defined, which provides reasonable pointers as to what sources are to be used and what quantity is needed.

4. Varying Integrity of data sources could be a big problem especial as data objects interact with each other. Taint propagation theory best explains this problem.

…CONCLUSION

5. Social media bias will remain a problem as making mistakes and lying are known bad issues with humans. If it is the biggest source of “fake news” then it’s intergrity quality should be questionable.[10]

6. Data generated from sensors and crowdsensing share similar problems. Multiple factors play key roles in creating and transmitting false or inaccurate data.

There is no perfect data anywhere.

Assigning weights to each data source based on some measure of data quality can go a long way in reduce the unwanted problems associated with each source.

These weights can also be varied and observations made so as to get a sweet spot.

THANK YOU!

REFERENCE

[1] D. Agrawal et al., “Challenges and Opportunities with Big Data,” Cyber Cent. Tech. Rep., p. 19, Jan. 2011

[2] Danah Boyd and K. Crawford, “CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon,” Inf. Commun. Soc., vol. 15, no. 5, pp. 662–679, Jun. 2012.

[3] D. Bollier, Communications and Society Program (Aspen Institute), and Aspen Institute Roundtable on Information Technology, The promise and peril of big data. Washington, DC: Aspen Institute, Communications and Society Program, 2010.

[4] D. Yang, G. Xue, X. Fang, and J. Tang, “Incentive Mechanisms for Crowdsensing: Crowdsourcing With Smartphones,” IEEEACM Trans. Netw., vol. 24, no. 3, pp. 1732–1744, Jun. 2016.

[5] F. Morstatter, “Detecting and mitigating bias in social media,” in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 2016, pp. 1347–1348

[6] J. Heikenfeld et al., “Wearable sensors: modalities, challenges, and prospects,” Lab. Chip, vol. 18, no. 2, pp. 217–248, 2018

[7] M. Dapkus and J. Pridotkienė, “Testing of Macroeconomic Lying Hypothesis: French Industry Case,” Procedia - Soc. Behav. Sci., vol. 156, pp. 286–291, Nov. 2014

[8] M. Rimando et al., “Data Collection Challenges and Recommendations for Early Career Researchers,” He Qual. Rep., vol. 20, no. 12, p. 14, Dec. 2015

[9] R. Ganti, F. Ye, and H. Lei, “Mobile crowdsensing: current state and future challenges,” IEEE Commun. Mag., vol. 49, no. 11, pp. 32–39, Nov. 2011

[10] S. Stephens-Davidowitz, Everybody lies: big data, new data, and what the internet can tell us about who we really are, First edition. New York, NY: Dey St, 2017.