How to Value Public Science Employing Social Big Data?

Maria L. Loureiro and Maria Alló

Contents

1	Introduction	93
2	Data: Twitter Data Collection	94
3	Data Visualization: Word Clouds	95
4	Results from Emotion Analysis and Hedonometer	95
5	Conclusions	97
	References	98

1 Introduction

Scientific discoveries can be classified as public goods. Arrow [1] discussed properties of knowledge that make it a public good; highlighting in particular, the fact that it cannot be depleted when shared, and once it is made public others cannot easily be excluded from its use. So, public good is a commodity or service that is provided without profit to all members of a society, either by the government or by a private individual or organization. Thus, a global public good is a public good that goes beyond borders, and CERN scientific output is the perfect example of a global public good.

A crucial issue for the supply of such global public goods is based on their value. When analyzing scientific contributions, it is important to note that there are two kinds of values: use values and non-use values. When talking about science we can identify the use value in the case of patents, licenses, and other market realizations of value. The non-use value is derived from potential market realizations that may be

M. L. Loureiro (✉) · M. Alló
University Santiago de Compostela, A Coruña, Spain
e-mail: maria.loureiro@usc.es

achieved based on scientific discoveries. The category of non-use values contains those denoted as option values, the bequest values and the existence values. The option value refers to the value given to a resource or service that is of no use today, but maybe extremely valuable in the future. In the case of the bequest value, this denotes the possibility of transmitting knowledge and cultural heritage to future generations; while, the existence value is the willingness to pay to preserve a resource or service because of its mere existence and not necessary because any use or benefit can be derived from it. Therefore, the Total Economic Value (TEV) of science should be considered estimating the sum of the three previous types of values. This a complex estimation and somewhat unknown in the short term. Even in the long run, the TEV can also be difficult to compute in the long term due to the existence of risk and uncertainty.

Thus, going back to the initial question, how can we value public goods? It is important to note that in some cases there are no market prices, therefore one option could be to ask citizens their willingness to pay. But, it is also important to note that the information is crucial and, for example, the participation and perceptions of stakeholders is fundamental. In this respect, we can adopt a direct approach asking citizens about their preferences (where different biases may arise) or an indirect approach where we obtain information from the market to obtain insights about how much we value science.

The aim of this work is to analyze the capacity of big data sources to evaluate the TEV of science; focusing specifically on non-use values. Einav and Levin [2] described how much economic research has evolved in the area of big data and new private datasets, showing a significant amount of questions can be addressed now. As an example, this paper provides an analysis about how we can measure the perceptions that citizens have about CERN through the information collected from the social media Twitter.

2 Data: Twitter Data Collection

Social media can be extremely useful to value global public goods, they provide us unsolicited opinions about current projects and is open data as in the case of Twitter. Specifically, Twitter is a social network that was born in 2006 and the place where we can interact with people around the world. These messages are known as “tweets” and they are limited to 140 characters. According to the data of 2019, Twitter has around 261 million monthly active users worldwide.

Data collection was conducted during October, November, December 2018 and from February–June 2019. This process has been based in the search of tweets through the use of hashtags and keywords. In order to collect the data, requests have been made through the library Tweepy for Python which works with Twitter Streaming API specifying a “keyword” or specific “hashtag”. At the same time that tweets were received, another request was made to obtain some additional data regarding the publication of the message. In addition, we were interested in gaining information about the language used. Specifically, the following hashtags were used to download tweets: @CERN, @AtlasExperiment, @LHC News, @CERN-LHC

Live, @ALICE Experiment, @CMS Experiment. On average, during this period, we recorded around 698 tweets in English and 247 in Spanish.

3 Data Visualization: Word Clouds

World clouds are effective ways of showing the most predominant topics in a text. They are frequently used for blogging and micro-posting. In particular, Fig. 1 shows the most relevant topics related to climate change in both countries.

In terms of the predominant words, we observe that Tweeter conversations in Spain are related to “cern”, “code”, “open”, “Microsoft”, “software”. In the case of English conversations, words related to “cern”, “lhc”, “collider”, “hadron” are the most predominant.

4 Results from Emotion Analysis and Hedonometer

4.1 Analysis of Emotions

In order to analyze our social media dataset from Twitter the lexicon developed by the National Research Center Canada (NRC) is employed; specifically, the dataset EmoLex [3]. The authors focused on eight different emotions: joy, sadness, anger, fear, trust, disgust, surprise and anticipation [4] and the type of sentiment being positive or negative. The lexicon employed contains around 14,000 words and

Fig. 1 Word clouds from tweets in English and Spanish

Graph 1 Analysis of emotions: English speakers

Graph 2 Analysis of emotions: Spanish speakers

25,000 word senses. As indicated, they identified a list of words and phrases and used Amazon’s Mechanical Turk to obtain emotions annotations.

The two graphs below show that out of the eight emotions evaluated, the most popular one in both English and Spanish tweets are “trust”, and “anticipation” (Graphs 1 and 2).

4.2 Hedonometer

In order to analyze the information retrieved from Twitter, the Hedonometer tool has been adopted. Previous studies conducted by Cody et al. [5] have also employed this

Table 1 Hedonometer results

How happy are you from 1 (unhappy) to 9 (very happy)?	
Spanish h_{ave}	5.545564
English h_{ave}	5.353653

technique. Specifically, the Hedonometer technique could be classified within the techniques of Natural Language Processing (NLP). It consists in the analysis of a text through its segmentation into phrases, and phrases, into words. Words are associated with scores of positive and negative feelings, and thus a total score for the sentence obtained, and then by aggregation, for the overall topic. In order to associate each word with a score, specifically, it uses the sentiment scores collected by Kloumann et al. [6] and Dodds et al. [7]. More specifically, the analysis is based on a lexicon or dictionary, which takes as reference, the 10,000 most used words in a long collection of articles from newspapers and books of Google Books as well as lyrics and a large number of generic tweets. These words have been classified according to the values of “happiness” or acceptance that 50 individuals have received from the Amazon Mechanical Turk platform.

The scale of happiness or acceptance has been measured on a Likert scale from 1 to 9, 1 being “least happy” and 9 counting as “most happy”, being able to segment the expression in “sadness or rejection” from 1 to 3, “indifference” from 4 to 6 and “happy” or “acceptance” from 7 to 9. The average happiness of users has been calculated following Eq (1):

$$h_{avg}(T) = \frac{\sum_{i=1}^N h_{avg}(w_i) f_i}{\sum_{i=1}^N f_i} = \sum_{i=1}^N h_{avg}(w_i) p_i, \quad (1)$$

Where T refers to a text, f_i is the frequency of the i th word w_i for which we have an estimate of average happiness, $h_{avg}(w_i)$, and $p_i = f_i / \sum_{j=1}^N f_j$ is the corresponding normalized frequency. From all these words, we have previously excluded the “stop words” (or words that are necessary for the grammatical construction but that by themselves have no meaning).

Results obtained are presented in Table 1. As it can be seen, the average happiness for both, English and Spanish speakers when talking about CERN is 5.54 and 5.35 respectively.

5 Conclusions

The above results show, accounting for all content published during the period of download, an overall positive feeling towards CERN research activity; and particularly this is reflected by the Spanish speakers. Such positive outlook may support a willingness to pay for CERN scientific services that should be further investigated.

References

1. Arrow, KJ. (1962). The economic implications of learning by doing. *The Review of Economic Studies*. Oxford Journals. 29 (3): 155–73. doi:<https://doi.org/10.2307/2295952>.
2. Einav, L. and Levin, J. (2014). Economics in the Age of Big Data. *Science*, 346, 715–721. <https://doi.org/10.1126/science.1243089>
3. Mohammad, SM., Turney, PD. (2013). Crowdsourcing a Word–Emotion Association Lexicon, *Computational Intelligence*, 29 (3), 436–465.
4. Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience* 1, 3–33.
5. Cody, EM; Reagan, AJ; Mitchell, L; Dodds, PS; Danforth, CM. (2015). Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll. *PLoS ONE* 10(8): e0136092. <https://doi.org/10.1371/journal.pone.0136092>.
6. Kloumann, IM., Danforth, CM., Harris, KD., Bliss, CA., Dodds, PS. (2011) Positivity of the English language. Available at <http://arxiv.org/abs/1108.5192> .
7. Dodds, PS; Harris, KD; Kloumann, IM; Bliss, CA; Danforth, CM.(2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, Vol 6, e26752.

Full Presentation

https://indico.cern.ch/event/727555/contributions/3461289/attachments/1868125/3073452/Valuing_The_Public_Good_dimension_of_Science-final.pdf

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.