Using textual data to measure Social Capital

Irene Saonara[1]

[1] Cognitive Science and Communication research Centre (CSCC) –
Università Cattolica del Sacro Cuore + Milano – Italy

Abstract

The aim of this paper is to investigate the possibility to develop a set of textual-data based indicators in order to use textual material such as interviews, national strategic plans and other official documents to evaluate the social impact of development projects. One of the most difficult part in monitoring and evaluating social impact is data collection and elaboration. In this perspective, the possibility to use some textual data from formerly available documents such as interviews, official documents and reports in order to evaluate social impact could be very interesting not only for researchers but also for public sector and NGOs' officers. Firstly, we decided to test our hypothesis on a small corpus, which could be analyzed using both qualitative and quantitative methods. There are not universally-accepted proxies for Social Capital presence and for this reason we chose to classify manually using our interviews in order to check the reliability of our method.

Keywords: Textual data, Social impact, Mixed methods

1. Introduction

The purpose of this paper is to investigate the possibility to develop a set of textual-data based indicators in order to use textual material to evaluate the social impact of development projects. An extensive discussion has been set up about the possibility to define and measure social impact and its main output, Social Capital. For the purposes of this essay, the term Social Capital indicates institutions, relationships and norms that improve the quality and quantity of society's social interactions and facilitate collaboration among actors. Using textual data to measure the social impact of a policy appears ambitious and unusual. However, this hypothesis is quite intriguing, considering the amount of textual data that are available through surveys and advocacy publlications. The main question that motivates this study is to find out the potential of textual quantitative methods for a deeper understanding of the complex phenomena characterized by a great number of possible proxies, instead of developing a general social impact monitoring and evaluation methodology based on textual data. This paper provides an overview of methodological issues and choices faced during the preliminary work for the "Love Matters in Policy Making: The Stop TB partnering process", a research project of World Health Organization (WHO) TB office and CSCC focused on exploring the transformative potential of community-based health policies, both at the individual and at the societal levels.

2. Corpus and Methods

2.1. Data

Our case study is the initiative *Stop Tuberculosis (TB) Partnership*[1] (Stop TB), implemented by the World Health Organization (WHO) from 2010 to 2014. Our corpus includes 13 interviews referring to the Stop TB Partnership experience. The interviews were collected in

[1] This project is now included in the WHO *End TB strategy* and its aim is to promote national partnerships against TB, involving public health organizations, private sector and NGO's. The website of the project is http://www.stoptb.org/

2012 by the central Stop TB Partnership Secretariat on the basis of an interview guide. All the interviews were recorded and fully transcribed by the same WHO officer. The whole corpus mirrors a vocabulary of V = 2 324 types and includes N = 22 221 tokens. The Type-Token Ratio (V/N) amounts to 10.5% (<20%), and the percentage of words with only one occurrence (hapax legomena) is 41.2% (<50 %). The corpus was pre-processed by means of the software Taltac2 (Bolasco et al., 2010) and Wordstat (Provalis Research, 1998), the textual data were processed by Iramuteq (Ratinaud 2009; Retinaud and De´jean 2009). By considering further criteria which are relevant to assess corpora homogeneity suggested by (Berruto, 1995), our corpus respects language variation (all interviews are in English, even though from non-native speakers), as well as diaphasic (all texts are transcriptions of spoken conversations), diachronic (all interviews took place in the same period, 2012), and diamesic variation (they are all spoken-transcripted texts) were offset. Diatopic and diastratic variation are quite large (in terms of social class, educational level and nationality), but considering interviews made to people who play the same role (leaders and focal points of the partnership) in their partnership programme was essential for our purpose, despite their different background.

2.2. Quantitative Analysis

Considering the lack of homogeneity of our corpus *hapax legomena* and words that occurred just in one interview were excluded from the vocabulary. Firstly, we considered most frequent content words in the whole corpus, then we computed the Term Frequency Inverse Document Frequency (TFIDF) index for each word. After this first analysis, considering the possible differences between our speakers we decide to check also the content words included in all interviews (*Common words*. Considering also the complexity of the concepts we are looking for, we found that analyzing simple words frequency, weighted frequency and distribution could not be enough. For this reason, the pre-process of the corpus was performed with Taltac2, in order to detect some meaningful units formed by two or more words, which could be useful to recognize Social Capital components. Using these words and meaningful units that could be considered as good proxies of Social Capital, we built our Social Capital vocabulary. The keywords' frequencies in the whole corpus and their distributions among different interviews were computed. In order to get a better representation of the complex phenomenon of Social Capital and its positive outputs we constructed some semantic fields mainly based on the keywords and on our definition of Social Capital, and then we considered their distribution. Our categories are *Working Together* and *Improvement*.

2.3. Qualitative Analysis

The quantitative analysis was built considering the suggestions contained in (Cresswell and Plano Clark, 2007) on mixed-method research. In this case, the aim of our qualitative analysis was just to detect if the quantitative results were reliable. For this reason we categorized every sentence of our corpus using an ex ante methodology, based on two questions, which are:

- Do our speakers tell us different stories or they just use different words?
- Are semantic fields the right tool to measure Social Capital?

This led to create two categories (Description of Partnership experience and Social Capital Elements) and 9 codes. Moreover, we created a Satisfaction sentiment based indicator to separate positive interviews from negatives. This classification is based on a comprehensive analysis of single interviews, and is not free of subjectivity of researcher issue.

3. Results and Conclusions

3.1 Different words, different stories?

Interestingly, even the distribution of the few common words selected as keywords is really not uniform. The lack of homogeneity of our corpus led us to wonder if these differences our speakers are telling us different stories or same stories but using different words.

3.2 Are the semantic fields the right tool?

Considering semantic fields' distributions we find out that there are some countries (such as Kenya, Thailand Pakistan) with no or just one corresponding result. As said before, this fact should not be interpreted as absence of Social Capital elements in Kenya's and Pakistan's stories. This fact only confirm us the necessity to consider more aspects, more words and keywords. Forming a semantic field requires to balance both lexical richness and semantic affinity. You have also to reduce as for as possible meaning ambiguity and duplicity. In this case, we decide to include in our semantic fields only words and expressions strictly related to our definition of Social Capital and used to describe an improvement in actors' capacity building or collaboration among actors, that occurs at least in two interviews and that are quite clear and not ambiguous even without considering them in their context.

In example, we include ABLE TO MOBILIZE and ABLE TO SHARE but we exclude ABLE TO REACH because for the first two we can determine that they are referred to Social Capital concept even without considering the context, while the third one could be referred also to sanitary target. The most important evidence is that most appreciated changes related with the partnership experience were those connected with an increase in ability and capacity building. Forming the Working together semantic field (Figure 1) our firs aim was to include all expressions that could be related with collaboration and sharing aspects of Social Capital. In addition, in this case, as in the improvement, we choose a mixed approach and we decide to consider firstly working together verbs (SHARE, MOBILIZE, JOIN, COORDINATE) and their derived but also some figurative verbs (BROUGHT TOGETHER, CREATE SYNERGIES) and some other expressions that include the adjectives SOCIAL or COMMON. This fact could led us to hope that our semantic camp is appropriately formed. Second, every bar include at least one expression related to JOINT and derived, one related to COMMON and GROUP or to SHARE and derived. This result suggest that even in the interviews characterized since now from a limited presence of Social Capital elements, such as Kenya or Pakistan, tell us a story of collaboration and sharing.

3.3 Qualitative validation

As can be seen in Figure 2 and Table 2 qualitative analysis partially validates quantitative results. If we consider Kenya and Thailand interviews we can see that red bars are really small in figure 4 and *Satisfaction sentiment based indicator* is 1. On the other hand Pakistan, which seems characterized by a low presence of Social capital keywords in Figure 1, contains a lot of sentences coded with Social Capital Codes (690 word on the total of 1349) and it's speakers is very positive about the partnership experience. Also if we consider countries with a lot of Social Capital keywords such as Swaziland and India we can see that their red bars are smaller and in India case the satisfaction *Sentiment based indicator* is just 2.

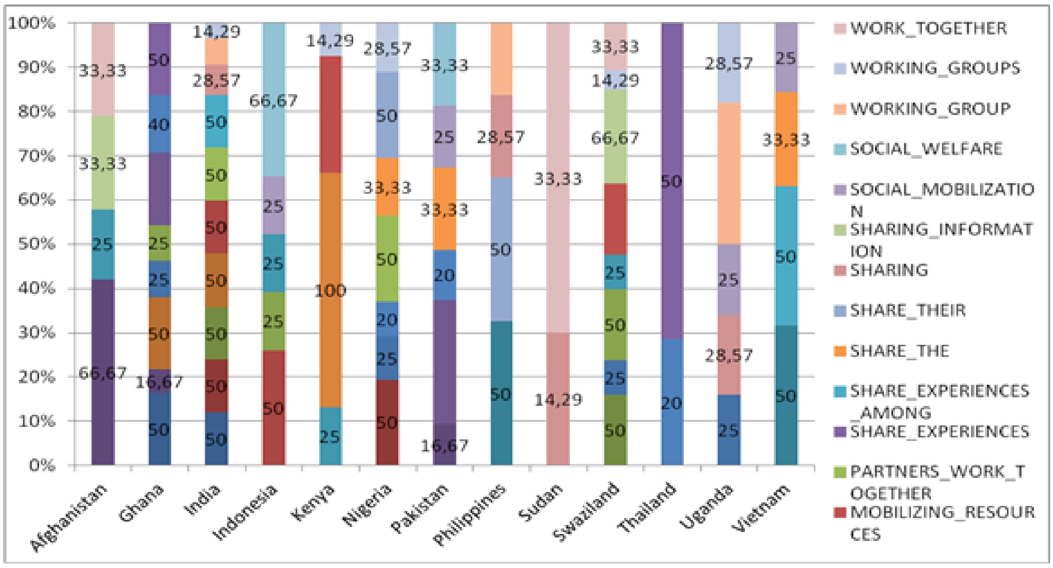

Figure 1: Distribution of words included in the semantic field Working together. Percentage are related to total occurrences of single expressions.

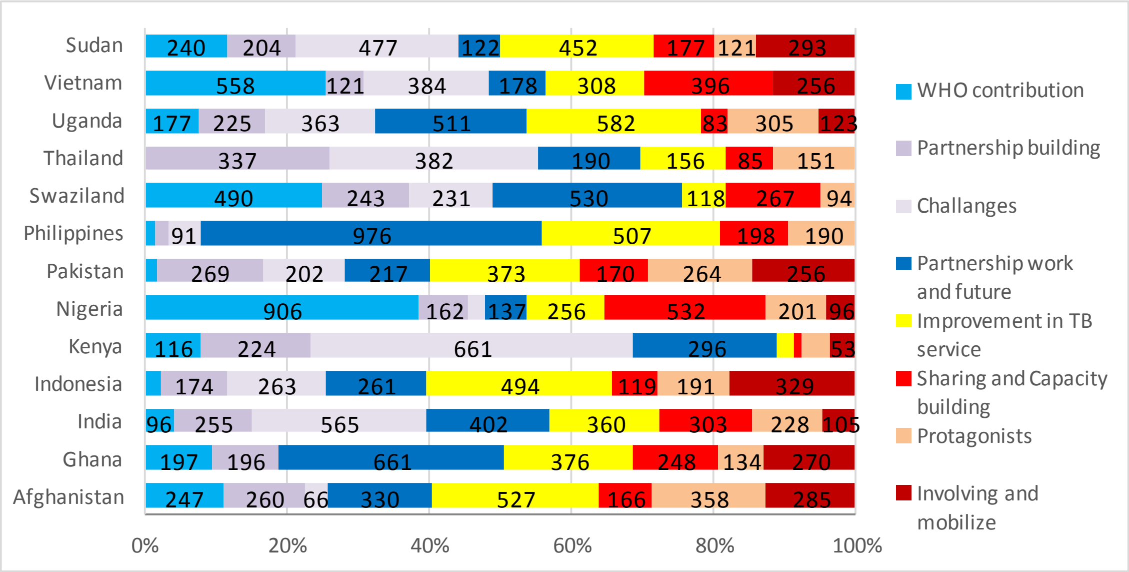

Figure 2: Words count per codes. Percentage are related to total word of single interview. We underline descriptions codes in blue shades and Social capital codes in red shades. Bars related to Improvement in TB service were colored in yellow to facilitate the difference between sanitary improvements (included in partnership experience but not related to social impact) and Social capital related codes.

Satisfaction sentiment based indicator		Countries
1	Negative expressions, low hope and trust in future improvements, mainly describe problems and fund issue	Indonesia, Kenya, Thailand, Philippines
2	Describe partnership building and underline some improvements but mainly focus on fund issue and difficulties. Contains explicit reference to non-national actors as main protagonists.	India, Uganda, Vietnam
3	State good results, underline working together experience and consequent improvements, trust and hope.	Afghanistan, Ghana, Nigeria, Swaziland, Sudan, Pakistan

Table 1 Description of Sentiment based indicator

The findings of this study provide some useful insights for the application of textual analysis techniques to evaluate the social impact of projects and policies. The differences between quantitative and qualitative results (e.g. the case of Vietnam or Indonesia) suggest that using just quantitative techniques might not be enough to map complex phenomena and concepts such as Social Capital. On the other hand, considering the case of Pakistan we can see that if the semantic camp is rather large and comprehensive (in our example *Working together*) quantitative and qualitative results coincide. In conclusion, our results must be interpreted with caution because of their dependence on the semantic fields composition. This research has thrown up many questions in need of further investigation. Additional analysis should focus on improving and testing semantic fields, in order to obtain more comprehensive "bags of words". It would be also interesting to compare our results with other obtained by available measures and textual tecniques' output such as Reinert's method (Reinert, 1990) or Labbé's intertextual distance (Labbé, 2001). Furthermore, more research is required to determine the efficacy of the presence of Social Capital components in the interviews as an indicator of the real outcome of partnership experience. Notwithstanding these limitations, the study suggests that using textual data to measure Social Capital could led to a deeper comprehension of this phenomenon, based on formerly available data.

Acknowledgements

I would like to express my deep appreciation to Professor Ajuna Tuzzi (FISSPA, Università degli studi di Padova), Professor Erika Teodora Uberti (CSCC, Università Cattolica del Sacro Cuore) and Dr. Stefano Sbalchiero (SPGI, Università degli Studi di Padova) for their valuable and constructive suggestions during the planning and development of this research work.

References

Berruto G. (1995). *Fondamenti di sociolinguistica*. Editori Laterza

Bourdieu P. (1980). Le capital social. *Actes de la recherche en sciences sociales*. vol.(31): 25-30.

Coleman J.S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*. vol.(5):35-42.

Cresswell J.W. and Plano Clark V.I. (2007). *Designing and conducting mixed methods research*. Sage Pubblications

Labbé C. and Labbé. D. (2001). Inter-textual distance and authorship attribution. *Quantitative Linguistic* vol.(8):213-231.

Popescu I., Macˇutek J. and Altmann G. (2009). *Aspects of Word Frequencies. Studies in Quantitative Linguistics*. RAM. Ludenscheid

Putnam R. D. (1995). Bowling Alone: America's Declining Social Capital. *Journal of Democracy*. vol.(6):65-78.

Ratinaud P. (2009) IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. www.iramuteq.org

Reinert M. (1990). Alceste. une methodologie d'analyse des donne´es textuelles et une application: aurelia de Gerard de Nerval. *Bulletin de Methodologie Sociologique*. vol.(26):24-54

Sbalchiero S. and Tuzzi A. (2015). Scientists' spirituality in scientists' words. Assessing and enriching the results of a qualitative analysis of indepth interviews by means of quantitative approaches. *Quality and Quantity*. vol.(04)1-16.

Woolcock. M. and Narayan D. (2000). Social capital: Implications for development theory. research. and policy. *The World Bank Research Observer*. vol. (15 nb.2).