Unsupervised spatial data mining for the development of future scenarios: a Covid-19 application

Yuri Calleo, Simone Di Zio

1. Introduction

In the framework of Future Studies, the development of future scenarios can contribute within the social context by providing inputs at the decision-making level in order to take action in the present. However, this implies an effort and a long-time frame in the first two phases of the scenario development (typically *Framing* and *Scanning*) which require a long desk research, such as reading of documents, research of the scientific literature or the consultation of experts for identifying the key factors (Bishop et al., 2007). In particular, the goal of the scanning phase is to define a number of basic driving forces which constitute the base for the construction of alternative futures scenarios. Some scholars (Kayser and Shala, 2020) estimated an average time of two weeks, in which typically the research team compose a panel aimed at understanding the object of study. Recently, with the exponential growth of social networks, users are constantly in connection with each other, disseminating textual, multimedia, and geographical content on a daily basis. It therefore follows that given the enormous increase in data sources within them and given the communication with which users share ideas, thoughts, and information, all this could be exploited in the context of scenario building.

From the premises made so far, we have developed a new approach that uses unsupervised classification models aimed at speeding up the first two phases of scenario development and optimizing the entire process. To capture the topics and the relevant key factors we used Machine Learning methods, including text-mining (Kayser and Blind, 2017) and Spatial Data Mining techniques. The goal of this work is to provide an answer to the following questions: "Is it possible to obtain information on the object of study by extracting key factors from Twitter?", "Does this approach speed up the Scanning phase?". And, above all, "What contribution can spatial data mining offer to the process of development of future scenarios?". To apply the method, we extracted a dataset from Twitter containing textual and geo-spatial content relating to Covid-19.

2. Materials and Methods

The approach used here applies unsupervised classification models belonging to Machine Learning and aims to extract the major topics within a dataset of tweets, in order to use them as key factors in the scenarios' development process. During the month of November 2020, a dataset of 60.000 tweets was extracted through the use of the Streaming API System using 95 keywords and hashtags related to the discussions on Covid-19 (Uhl and Schiebel, 2017). After extracted the matrix, we proceeded to import it into Python to clean and manipulate it, and then we applied the techniques useful for our analysis (after this phase, the remaining tweets resulted in 29.949). The first step carried out saw the conversion into numbers, better defined as "number vectors" (Atenstaedt, 2012) of the data matrix, through the "lemmatisation" and "tokenization". In the processing of a specific language, the vectors of numbers are determined by textual data, in order to reflect various linguistic properties of the text, where a coding of the characteristics is necessary (Goldberg, 2017). First of all, we tried to have a qualitative general view of our dataset by applying the text-mining technique using the bag-of-words model that extracts and flexibly represents the data of a given text describing

Yuri Calleo, Gabriele d'Annunzio University, Italy, yuricalleo@yahoo.com, 0000-0002-0190-6061
Simone Di Zio, Gabriele d'Annunzio University, Italy, s.dizio@unich.it, 0000-0002-9139-1451

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Yuri Calleo, Simone Di Zio, *Unsupervised spatial data mining for the development of future scenarios: a Covid-19 application*, pp. 173-178, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.33, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

the occurrence of words within a document or corpus of documents. The same extracts in a document only the words known and therefore present in a vocabulary assigned to it, while any other information is discarded a priori. We then applied a Sentiment Analysis to understand the degree of polarity of the terms found within the dataset, using two distinct algorithms, called *Vader* and *Afinn,* in order to have a comparison between the two results obtained. We decided to use these algorithms since they are two of the most used in Sentiment Analysis for social networks (cfr. Narasamma et al. 2021; Mayor & Bietti, 2021; Tan & Guann, 2021), and compared to the others, they are able to specifically decipher abbreviations and emojis in the corpus of documents.

They are tools based on lexical rules in relation to what is published mainly on social networks (Hutto & Gilbert, 2014) using a vocabulary of words generally labelled a priori (manually) and subsequently acquired by the model based on their semantic orientation (for example they can be labelled as positive, negative, or neutral). Both algorithms assign a final score based on a sum of the valence scores of the terms in the text and normalized usually between the negative (-1) and the positive (+1) extremes (Huang et al. 2019).

After having a general view of the dataset, in order to understand the most cited terms and their polarity, we used topic modelling to extract possible topics and keywords from the tweets. In this case, we used the Latent Dirichlet Allocation (LDA) (Tong and Zhang, 2016) with the following term frequency function of term t:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ represents the raw count of the term t in document d.

It is based on a distributive hypothesis of statistical measurement, through the extraction of a series of topics from a corpus of documents. This process is carried out through the mapping of every single document with a good part of the words present (Wang & Grimson, 2007), and the model assigns to each topic a word arrangement determining the key factors.

The LDA assumes that the topics follow a Dirichlet distribution (Minka, 2000), in fact the similarity of documents and topics is controlled by hyperparameters known as α and β; if α is low it will assign fewer topics to each document, while when α is high, we will have the opposite. A low β value will use fewer words in the topic modelling process, while a high value will use more words, thus making the topics more similar to each other. The LDA, in fact, does not know a priori the number of topics or terms to be extracted. The model produces a vector that contains the coverage of each topic for the document to be modelled: $\boldsymbol{c} = (c_1, c_2, \dots)$ where c_1 is the coverage of the first argument and so on.

To answer the research questions, we propose an analysis of georeferenced data that will optimize all process by adding important spatial information. Here we use the expression "georeferenced data" in a broad meaning, including any kind of information useful to link a tweet to a geographic object, where the object can be a unit of a vector shapefile layer (like for example a country). Numerous studies have been conducted on Twitter using text-mining or open-mining techniques (Pang & Lee 2008; Taboada et al., 2011; Liu 2012; Poria et al., 2014). Few studies, on the other hand, have focused on the construction of future scenarios starting from the extraction of georeferenced data from social networks. The spatial aspect, in our case, becomes of fundamental importance, as having a geographical view of the subject would benefit the development process.

In the scientific literature, some studies (including Haining, 2010) have highlighted the importance of georeferenced data and therefore the presence of such information in the data (if any) is worth to explore. Actually, through web mining it is not easy to extract spatial information, given that the geographic coordinates (latitude and longitude) are rarely available. The social networks themselves, while previously freely providing quantities of data relating to the positions of users, recently they try to protect themselves by not having such data extracted in a substantial way. In our case, having used a streaming API system extraction, it was possible to model them.

First, we replaced the missing values with the wording "data_2 NA", after this first step, we have obtained 20.372 tweets with a geographic information included. Subsequently, we linked each tweet to the corresponding country from which it was written by means of information on the location (e.g. village or city), so that, for example, a tweet from Paris is assigned to France. From that, we obtained the frequency distribution of the number of tweets for each country, for each topic and for each key factor permitting to calculate the discussion rate in a single topic and in a single country. These relative frequencies are then reported in a GIS software (Q-GIS Development Team, Open Source Geospatial Foundation Project. http://qgis.osgeo.org.) to create cartograms for the topics, in order to have a representation of the spatial distributions of the same.

3. Results and discussion

The results obtained fully answer the research questions. In fact, it was possible, through the use of text-mining and spatial data mining techniques to extract the influencing factors from our dataset for future scenario development. From sentiment analysis it was possible to measure the polarity of the terms within our matrix, identifying more positive words than negative ones in both algorithms. In support of this analysis, the results are shown in Table 1.

Table 1: Sentiment Analysis

Algorithm	*Positive*	*Negative*	*Neutral*
Vader	18261	11688	51
Afinn	21119	8830	51

The key factors were extracted through topic modelling which highlighted 5 topics with 6 related keywords (Table 2). The first topic focuses on the *health* aspects, important to understand how the existing pandemic has brought health problems, causing discomfort and death. But beyond the physical aspect, the psychological aspect has also been affected, in fact we can find the presence of the term "anxiety" within the keywords that compose our topic. However, the vaccination uncertainty that persisted in November 2020 should not be underestimated, this aspect is of fundamental importance precisely because it fuelled – and feeds – discussions and conspiracy theories (see topic 3). The second topic describes the *political* aspects, and it is worth noting that the dataset, having English as language and having been extracted during the American elections, was affected by a strong influence of the same. This perspective can be observed from the terms: "government" and "trump". The third topic is reserved for denial and *conspiracy*, as we can see from the words "forced", "reality", "protest" and "planning".

Table 2: Topic modelling

Topic	*key1*	*key2*	*key3*	*key4*	*key5*	*key6*
Health	covid	quarantine	death	pandemic	vaccine	anxiety
Politics	politics	government	trump	gates	underperformance	progressivism
Conspiracy	forced	talking	reality	protest	blaming	planning
Economy	economy	bottomed	lockdown	employee	million	recession
Society	social	distancing	app	people	black	track

The fourth topic refers us to the *economic* field, in fact governments (see keyword "recession") were suffering for the pandemic. Citizens too (see keyword "employee") – forced to close shops, companies, etc. to prevent the spread of infections, they found themselves in a difficult grip to overcome with consequences at work and personal level. Finally, the fifth topic regards the *social* context, and shows how the pandemic issue has had implications in the social structure. Social distancing adopted by governments prevented the normal development of social activities. Not only that, the spread by governments of applications aimed at tracking movements has also had a debate on social networks, specifically on possible complications and on the possible violation of citizens'

privacy. The results are shown in Table 2. After analysing the keywords for each topic, we constructed a cartogram for each of them (Figures 1-5).

Figure 1: Health

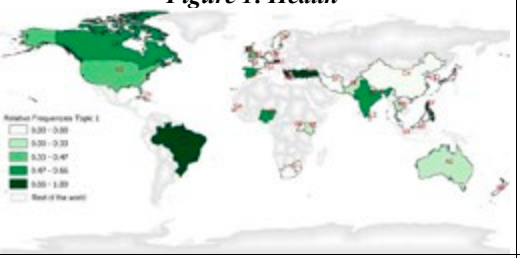


Figure 2: Politics

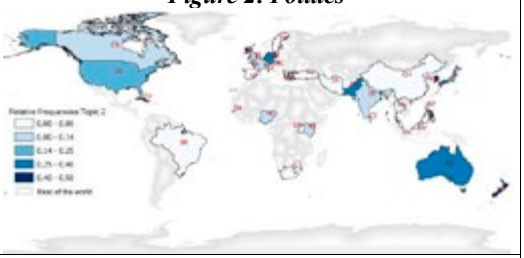


Figure 3: Conspiracy

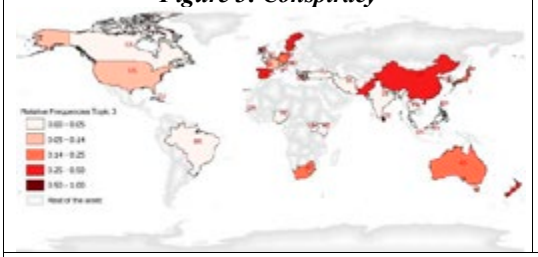


Figure 4: Economy

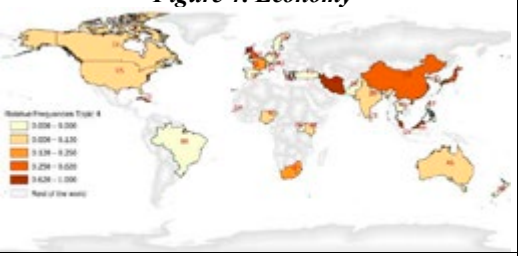
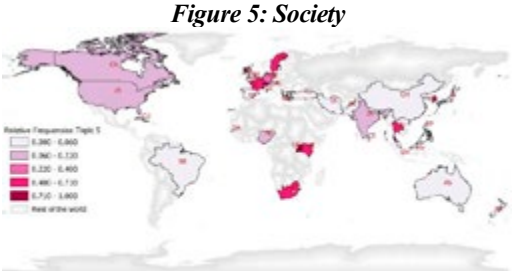


Figure 5: Society



Specifically, topic 1 (fig. 1), concerning health, it was carried out in Austria, Brazil, Canada, Greece, Philippines and Turkey. These rates of discussion may be higher than in other countries because during the period studied these countries were experiencing more infections and deaths from Covid-19. As for topic 2 (fig. 2) which analyzed political discussions on Twitter, it was more discussed in American countries, Australia, Germany, South Korea and New Zeland, probably due to the political involvement of the american elections carried out in November 2020. Topic 3 (fig. 3) – which analyzed the conspiracy aspect – sees a multitude of countries involved, take for example Spain, Sri Lanka, New Zeland, China and Pakistan. China, first saw the virus appear in its territory, and subsequently it had to interface with conspiracy theories about the nature of the virus trying to disprove them rapidly. Topic 4 (fig. 4), depicting the discussion rates of the economic topic, was most discussed in Iran, China, Japan, Malaysia and United Kingdom, probably because they were particularly affected by the economic damage that has occurred resulting in a strong response from central governments. The last topic (fig. 5), depicting the discussion of topics of a social nature, finds its foundation in Singapore, Switzerland, Uganda and Sweden. A specific note must be addressed in the analysis of African territories, in fact it is possible to find a strong rate of discussion in some countries such as Nigeria, Uganda, Gambia and Kenya compared to other continents, probably due to the social problems added by the pandemic issue to those already existing.

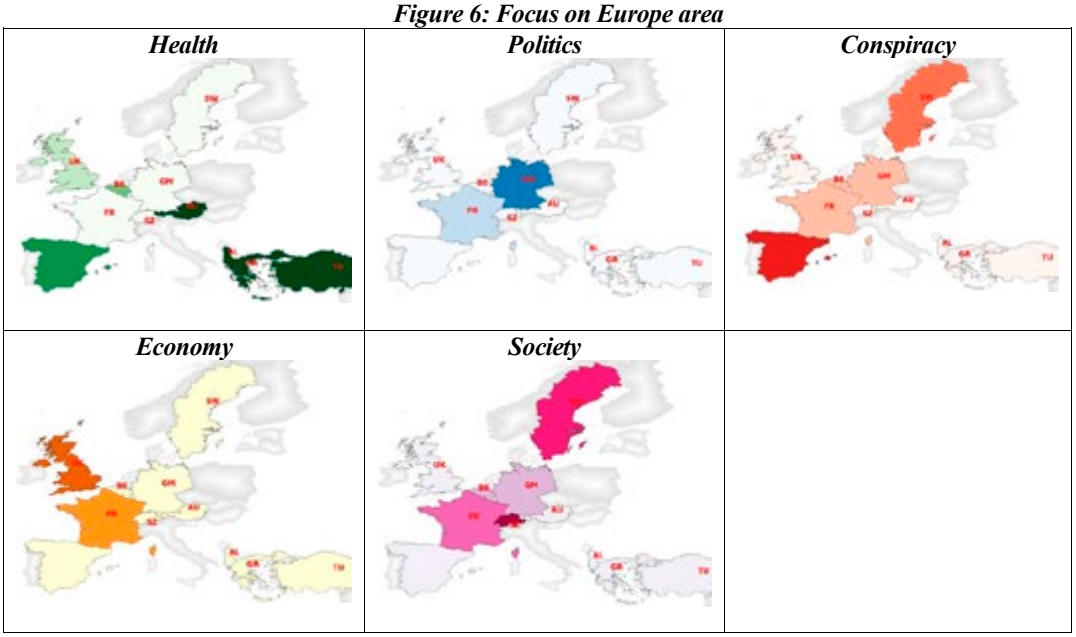
Since the world scale does not allow to highlight all the details, especially for the smaller countries, in Figure 6 we report the five cartograms with a focus on the European region.

The analysis of georeferenced data has fully answered our research questions, given that the results can be used in the context of futures studies in order to implement the initial process of

constructing futures scenarios. This approach provides an effective tool for the development of future scenarios greatly reducing the timing of the *Framing* and *Scanning* phases. Furthermore, it provides a contribution to the future research from a statistical-spatial field and, in particular, in the field of spatial scenarios.

Starting from these results, the scenario planning process will continue with the forecasting phase (Bishop et al., 2007; Hines and Bishop, 2015), which consists of the generation of a sufficient number of alternative futures.

Figure 6: Focus on Europe area



4. Concluding remarks

The approach developed above confirms the possibility of introducing the use of text-mining and spatial data mining within the first two phases of the scenario development (Framing and Scanning). It was therefore possible to extract the influencing factors in a short time frame without any literature review of the object studied and without the consultation of experts. Our study, in addition to providing elements for speeding up the process, enrich the analysis through the spatial component that offers important insights, when it is possible to observe the dynamics on geographical distributions. Understanding in which situations and in which parts of the globe a certain key factor is spoken of, means that much more information is provided. The analysis of Twitter data is only a starting point, in fact, in future studies additional social networks could also be taken into consideration (e.g., Reddit, Facebook, Instagram etc.). Furthermore, it will be possible to analyse much larger datasets in order to have a more complete vision of a given subject.

We recommend that subsequent studies focus on the spatial analysis, too often underestimated in futures studies, but capable of providing important information and, if combined with text-mining techniques, it could lead to an important turning point in the process of scenario and/or spatial scenario development.

It is worth noting that the method proposed in this paper produces spatial data that can be analyzed with the typical tools of spatial statistics. For example, a spatial autocorrelation analysis could reveal similarities between adjacent countries, even if in this case study it was not possible given the very low contiguity of the nations included in the dataset.

References

Atenstaedt, R. (2012). Word cloud analysis of the BJGP. *British Journal of General Practice*, **62**(596), pp. 148-148.

Bishop P., Hines A., Collins T. (2007). The current state of scenario development: An overview of techniques, *Foresight*, **9**(1), pp. 5–25.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, **10**(1), pp. 1-309.

Haining, R. P. (2010). The nature of georeferenced data. Handbook of applied spatial analysis. *Springer*, Berlin, Heidelberg, pp. 197-217.

Hines A., Bishop P., (2015). *Thinking about the Future: Guidelines for Strategic Foresight*, 2nd Edition, Hinesight Edition, Huston (TX).

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, **167**, pp. 26-37.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, **8**(1).

Kayser, V., & Blind, K. (2017). Extending the knowledge base of foresight: The contribution of text mining. *Technological Forecasting and Social Change*, **116**, pp. 208-215.

Kayser, V., & Shala, E. (2020). Scenario development using web mining for outlining technology futures. *Technological Forecasting and Social Change*, **156**, 120086.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, **5**(1), pp. 1-167.

Mayor, E., & Bietti, L. M. (2021). Twitter, time and emotions. *Royal Society open science*, **8**(5), 201900.

Minka, T. (2000). *Estimating a Dirichlet Distribution*. MIT Technical Report, Cambridge, (US).

Narasamma, V. L., Sreedevi, M., & Kumar, G. V. (2021). Tweet Data Analysis on COVID-19 Outbreak. *Smart Technologies in Data Science and Communication*, Springer, pp. 183-193.

Pang, B., & Lee, L. (2008). Using very simple statistics for review search: An exploration. In Coling 2008. *Companion volume*: Posters, pp. 75-78.

Poria, S., Cambria, E., Winterstein, G., & Huang, G. B. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, **69**, pp. 45-63.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, **37**(2), pp. 267-307.

Tan, M. J., & Guan, C. (2021). Are people happier in locations of high property value? Spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data. *Applied Geography*, **132**, 102474.

Tong, Z. and Zhang, H., (2016). May. A text mining research based on LDA topic modelling. In *International Conference on Computer Science, Engineering and Information Technology*, pp. 201-210.

Uhl, A., Kolleck, N. and Schiebel, E., (2017). Twitter data analysis as contribution to strategic foresight-The case of the EU Research Project "Foresight and Modelling for European Health Policy and Regulations" (FRESHER). *European Journal of Futures Research*, **5**(1), pp.1-16.

Wang, X., & Grimson, E. (2007). Spatial Latent Dirichlet Allocation. *NIPS*, **20**, pp. 1577-1584.