This is the peer reviewd version of the followng article:

Supervised Machine Learning Methods to Disclose Action and Information in “U.N. 2030 Agenda” Social Media Data / Sciandra, Andrea; Surian, Alessio; Finos, Livio. - In: SOCIAL INDICATORS RESEARCH. - ISSN 1573-0921. - 156:2-3(2021), pp. 689-699. [10.1007/s11205-020-02523-4]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

11/04/2024 01:48

(Article begins on next page)

Supervised Machine Learning Methods to Disclose Action and Information in “U.N. 2030 Agenda” Social Media Data

Andrea Sciandra, Alessio Surian and Livio Finos

Abstract In 2015, the United Nation General Assembly adopted the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals aiming at ending all forms of poverty, fighting inequalities, and tackling climate change. We collected Twitter data about the 2030 Agenda from May 9th to November 9th, 2018. The aim of this work is to obtain a classification of each tweet in the corpus according to the “Information” - “Action” categories, in order to detect whether a tweet refers to an event or it has only an informative-disclosure purpose. It seems particularly interesting to understand how and to what extent people and organizations are playing a more active role in shaping the process of responding locally and internationally to climate change. Explicit intention to act or inform had been captured by hand coding of a randomly selected sample of tweets and then the classification had been extended to the whole corpus through a supervised machine learning method. Overall, our classification supervised model has produced satisfactory results.

Keywords: supervised machine learning, textual data analysis, sustainable development goals, social media

1 Introduction

In 2015, the United Nation General Assembly adopted the 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs)[1] aiming at ending all forms of poverty, fighting inequalities, and tackling climate change. The Agenda is a plan of action including 169 targets, National agencies and monitoring actions. At the international level, core actors include the United Nations Development Programme, UNDP, and the United Nations Environment Programme, UNEP.

In the monitoring the 2030 Agenda for Sustainable Development it seems particularly relevant Ulrich Beck's conceptual framework for the study of local responses to global risks, and particularly climate change. Decentralized action is often studied in relation to national contexts. According to Beck et al. (2013) this approach falls short in relation to the global nature of this risk. Studies should be able to relate climate risk to the generative process encouraging the forming of cosmopolitan communities, i.e. facing a global risk should be mapped also in terms of cultural and political changes, and especially of those changes advocating a global perspective. Therefore, Beck et al. (2013) provide a theory concerning the formation of an operational capacity to face the climate change. Beck suggests the category of "emancipatory catastrophism" as a way to negotiate a third option in between the dichotomy that opposes an utopian approach, i.e. investing in "green" technology vs an apocalyptic approach, i.e. claiming a state of emergency and focusing on local conflicts. Emancipatory catastrophism implies a recognition of the limits of current knowledge and investing in a medium term scenario (the next two generations), halfway between the short term horizon of political urgency and the long term span of technical approaches (Groves, 2019).

[1] The 17 SDGs: 1. No Poverty, 2. Zero Hunger, 3. Good Health and Well-being, 4. Quality Education, 5. Gender Equality, 6. Clean Water and Sanitation, 7. Affordable and Clean Energy, 8. Decent Work and Economic Growth, 9. Industry, Innovation and Infrastructure, 10. Reduced Inequality, 11. Sustainable Cities and Communities, 12. Responsible Consumption and Production, 13. Climate Action, 14. Life Below Water, 15. Life on Land, 16. Peace and Justice Strong Institutions, 17. Partnerships to achieve the Goals.

As the 2030 Agenda is a plan of action, this paper intends to outline an interpretative schema by mapping the categories of information and action emerging in social media. It seems particularly interesting to understand how and to what extent people and organizations are playing a more active role in shaping the process of responding locally and internationally to climate change.

2 Data collection

Data about the 2030 Agenda had been collected through Twitter API using `twitteR` R package with Oauth authentication from May 9th to November 9th, 2018. By an access token associated to a Twitter app, we issued a search based on a text string (“agenda2030”). The resulting dataset for this study consists in a corpus of N = 209216 tweets. The Twitter streaming API allows users to gather up to 1% of all tweets that pass through the service at any time. Twitter users of course do not represent people as they are a particular sub-set and we cannot assume that accounts and users are equivalent (Boyd & Crawford, 2012). We did not look at the representativeness issue, because we were interested in empirically applying SML methods with human tagging, instead of providing a reliable distribution of the categories of information and action. However, some authors have shown that these limits can be overcome by improving the sample population coverage (Sampson et al., 2015), e.g. by splitting the same keywords across multiple crawlers and, in wider terms, a social media analysis may capture population attitudes and behaviours, even if the characteristics of the users do not reflect the characteristics of the full population (Ceron et al., 2017).

We chose Twitter among the other social media because it is by far the most used platform for social science research, due to its API flexibility in terms of free access to data (McCormick et al., 2017). Our analysis covered a period of 6 months in 2019, with the aim of capturing reactions to: the first UN summit on the SDGs (New York, September 2019); the high-level political forum for follow-up and review of the 2030

Agenda (New York, July 2019); the preparatory work for the UN climate change conference Cop25 (Madrid, December 2019), which was to be held in Chile but had been moved to Spain. In this sense, we noted a particular interest on 2030 Agenda and a quantitative impact on Twitter of the Spanish language, specifically several tweets about Central-South America initiatives and later from Spain for the reasons mentioned above. We decided to focus on the Spanish language tweets, for specific research interests (Surian & Sciandra, 2019) and to achieve a certain degree of heterogeneity, including a large European country and most of Central and South America.

The increasing availability of digitized text, especially from Social Media, offers enormous opportunities for social scientists, even though this kind of data contain lot of noise. In fact, text fragments coming from social media can be considered as off-topic for the purpose of the analysis very frequently. It must be stressed that commonly off-topic texts don't use completely different words from the text of the training set, so classifiers will attribute as outcome this category with high probability and very rarely the true semantic category (Ceron et al., 2017).

3 Pre-processing

The first step in our analysis was pre-processing in order to remove noisy and inconsistent data and to improve classification effectiveness (Symeonidis et al., 2018). We applied the following pre-processing techniques before feature selection and text classification.

- Language control

We carried out a language detection, as we decided to focus on the Spanish language tweets (about 132000). Anyway, since we intend to use Supervised Machine Learning (SML) techniques, we must remember that these methods are completely independent of the language of the text; moreover, SML improves accuracy compared to a lexicon classification (Ceron et al., 2016) and there is no a priori set of categories.

- Retweet homogenization

We identified and homogenized any unmarked retweet by comparing texts with Levenshtein distance (RThound function - `TextWiller` R package). After these procedures, the retweets count is equal to 84.7%, a very high share.

- Text cleaning

In this step we meant to normalize text encoding and then removing URLs, emoticons, and punctuation. Therefore, all the URLs appearing in the corpus had been replaced with a tag (wwwurlwww). If an emoticon had been recognized, it had been replaced with a word that specify the kind of emoticon (e.g. ";)" had been recoded into "emote_wink").

- Dimensionality reduction

We removed stopwords and applied word stemming in Spanish language in order to remove noise (words that are not useful for classification) and to reduce dimensionality by merging terms with the same root. The initial corpus had 39193 types and 2726049 tokens (type/token ratio: 14%), while after pre-processing, removing stopwords and stemming we had 1482492 tokens and 35989 types (type/token ratio: 2%).

Figure 1 shows a word cloud of the most frequent words, excluding the search keyword "agenda2030", the tag for the URLs and the retweet tag.

4 Textual analysis

A textual analysis allowed us to include some variables related to SDGs hashtag and also the tf-idf weighting (Salton & Buckley, 1988), to show how important a word is to distinguish "Information" and "Action" tweets in our corpus. In particular, some simple text mining procedures allowed us to create the document-term matrix we used for the classifications. We worked within the "bag of words" framework, as only tf-idf weighting was applied to the words and we do not assume any Natural Language Processing rules nor do we want to use ontological dictionaries. We also analyzed the main multiword expressions by their frequencies as we tokenized adjacent words into

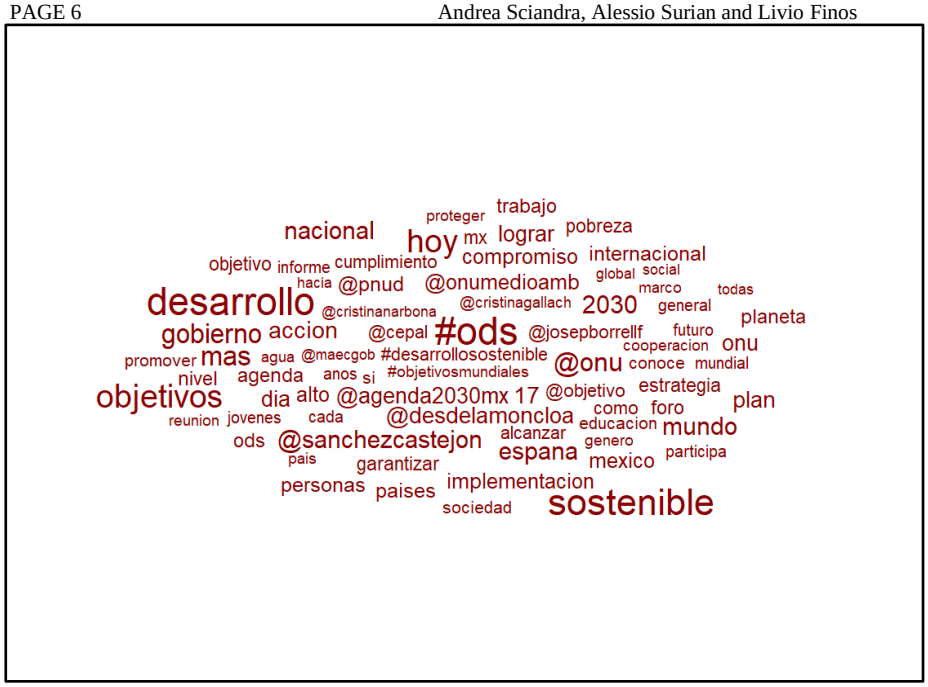

Figure 1: Most frequent words (wordcloud)

n-grams. The aim of this procedure was to identify some patterns in our corpus, potentially useful for the hand-coding phase, by visualizing some recurrent relationships among words simultaneously (Figure 2 includes the search keyword “agenda2030”, the tag for the URLs and the retweet tag). Moreover, in very few cases we recoded some n-grams because they were equivalent to acronyms (e.g.: “objetivos desarrollo sostenible” has been coded as “ods”, a very frequent acronym and hashtag).

Beyond the most frequent n-grams related to SDGs, two focal types emerged: “agenda2030” and “rt” (retweet). “Agenda2030” was associated frequently with an external link (URL) and with the 17 SDGs, while the retweets involved, among others, United Nation, Mexican party PNUD and Spanish Prime Minister Sànchez. Other frequent n-grams includes “high-level political forum”, “action plan”, “protect from poverty” and “civil society”.

Since we had a big sparse document-term matrix (35989 terms), we decided to allow maximal sparsity at 99.9%, or 0.01% in relation to document frequency. The

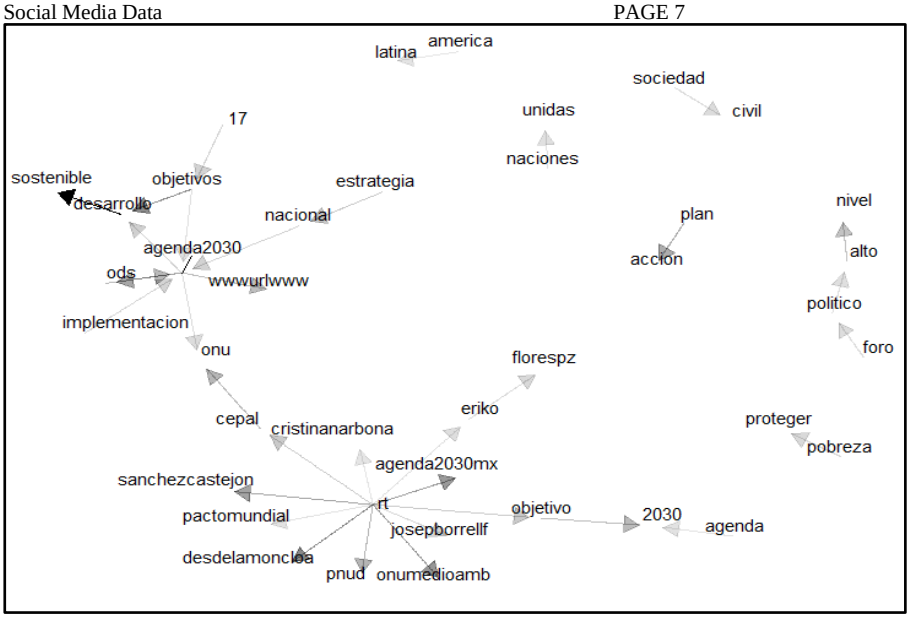

Figure 2: Most frequent bigrams

resulting matrix contains only 1745 terms, since we removed all the terms which have at least a 99.9% of empty elements (terms occurring 0 times in a document).

5 Training set

The aim of this work is to obtain a classification of each tweet in the corpus according to the “Information” - “Action” categories, in order to detect whether a tweet refers to an event or it has only an informative-disclosure purpose. Explicit intention to act or inform had been captured by hand coding of a randomly selected sample of 1584 tweets and then the classification had been extended to the whole corpus through a supervised machine learning method. The data had been hand coded by the three authors.

5.1 *Hand coding*

After randomly selecting a subset of 2000 (1584 unique), 2030 Agenda tweets were tagged manually in 2 (4) dimensions:

- "Information" (split into "General Information" or "Information with data").
- "Action" (split into "Action: Policies/Political Actors" or "Territorial Action").

The Information categories have been hand coded according to the following recommendations:

- General Information, i.e. tweets for merely informative purposes, which also include an invitation to follow a political event; this type is called "general information".
- "Information with data" to highlight the informative tweets where there are some data, or the data of a report are mentioned.

Otherwise, the dimension of the action should refer to creating something or at least a strong encouragement to do so. In particular, this category had been split into:

- "Action: Policies/Political Actors", i.e. tweets talking about national and local government as well as NGO actions or specific public policies.
- "Territorial Action": concrete projects and initiatives are cited, in other words, real actions on the territory. For "Territorial Action" a good indicator could be the word *taller* (laboratory/workshop), as well as *acción*.

The 4 dimensions are mutually exclusive and are distributed as follows:

- General Information: 72.35%
- Information with data: 10.48%
- Action: Policies/Political Actors: 9.66%
- Territorial Action: 7.51%

The agreement rate between coders was 90.06% and disagreement cases had been corrected following a discussion between the three coders.

6 Supervised Machine Learning

Among the many, we focused on Gradient Boosting method (Friedman, 2002, R implementation by Chen et al., 2019) for its flexibility and good performances. The Gradient Boosting method results to be particularly suitable for models characterized by sparse features since it produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

We chose to use all the terms of the reduced document-term matrix as features, weighted by their tf-idf. Some more variables collected through Twitter API were added as features (retweet count, favorite count, etc.).

The model is set to predict the four (or two) dimensions – i.e. classes – and is fitted to maximize the accuracy of the prediction. For the estimate of the proportion of each dimension – i.e. topic/class – we use the estimated probability of each tweet, so that each tweet contributes to the estimate of each dimension. We highlight that this is not standard for ML-based, which usually estimates the proportion of each dimension on the basis of classified documents.

7 Results

A summary of the performances of the model with four and two dimensions is reported in Table 1. Despite not being optimal, the indices demonstrate the ability to predict the dimension (i.e. the topic) of the tweet.

The importance of the most relevant features and words in the model (4-dimensions) are shown in Figure 3. The most important feature was the retweet count. This result did not surprise us from the statistical point of view, because this metadata provided by Twitter had a less sparse distribution compared to the selected terms and

Table 1 Performance indices of prediction models

Four Dimensions model			***Two Dimensions model***		
Accuracy	0.810		Accuracy	0.864	
	Precision	Recall		Precision	Recall
Info-Generic:	0.805	0.974	Info	0.869	0.988
Info-Data:	0.805	0.545			
Action-Politics:	0.750	0.300	Action	0.787	0.235
Territ.-Action:	1.000	0.217			

probably it might be a good feature to predict the largest category (General Information). Anyway, several words were among the most important features, e.g.: “laboratory”, “universal”, “municipalities”, “plan”, “rights”, “world”, “initiative”, “cities”, “project”, “council”, “poverty”. Some of these words are strongly linked to our categories, such as the already mentioned case of “laboratory” for territorial actions, and probably “initiative” for the category referring to political actions.

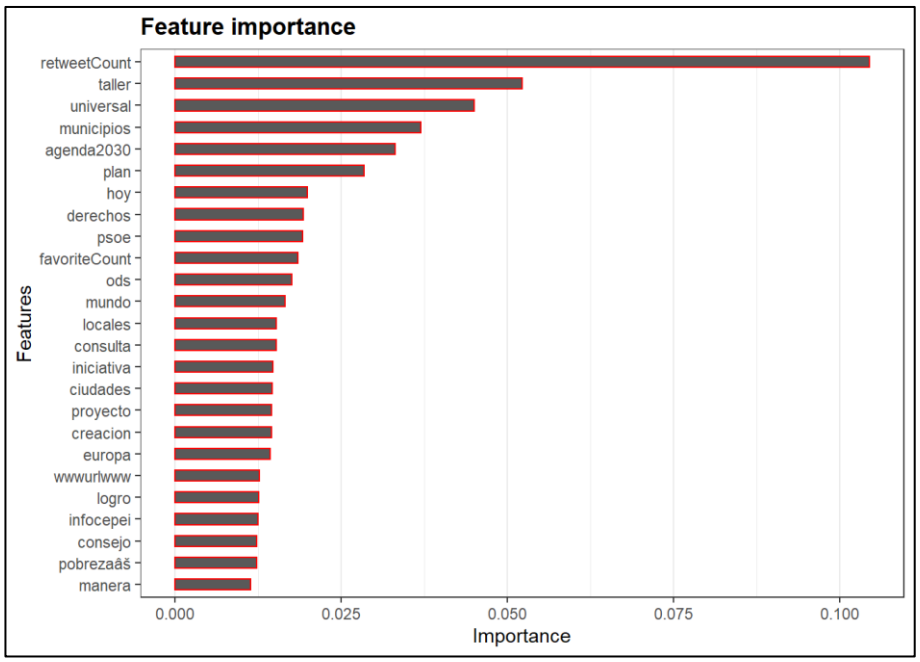

Figure 3: Importance of top 25 features

Table 2 Estimated percentage of the four dimensions (131029 documents)

Dimension	***Percentage***
General Information	75.74%
Information with data	8.74%
Action: Policies/Political Actors	9.16%
Territorial Action	6.36%

The estimate of dimension's proportion over the whole dataset (131029 documents) is given in Table 2.

8 Comparisons

The integrated Sentiment Analysis (iSA, Hopkins and King, 2010; Ceron et al., 2016) is one of the best-known approaches for the estimates of a set of categories in a corpus. iSA is based on a supervised hand-coding of a set of documents and it directly estimates the aggregated distribution of a given set of categories, through the frequencies of the terms of the corpus. So, this approach does not provide estimates for each single document, while it focuses on the estimates of the overall proportion of the categories in the dataset.

In this article we tackle the problem from a different perspective. Similarly to iSA, we start from our hand-coded training set, we use it as response variable in a supervised machine learning model, using the terms and other information as features. Here we run two experiments for the purpose of comparison among the two methods. Our hypothesis is that a well-trained machine learning method may be better than iSA when the frequencies of the categories in the train set (i.e. the part of the corpus that is hand-coded) are very different from the test set one (i.e. the not hand-coded). This could represent a typical situation of a phenomenon evolving over time, where the topics discussed tomorrow may be very different in proportion from the ones of the hand-coded dataset of today. This intuition arises from the fact that ML methods are

developed with the aim to exploit features to predict a topic and do not rely on the marginal distribution of topics on the training set (as iSA does). Therefore, they are more able to capture a change while extending the results from the training to the test set.

8.1 Balanced Experiment

In this first experiment we used the 1584 hand-coded documents as the whole corpus (all other documents are excluded here) and randomly sampled 500 of them as training set while the remaining are used as test set. We recorded the estimated proportions of topics for the whole dataset and we compared these estimates with the true proportions through the Mean Absolute Error (MAE). We repeated this procedure 100 times. Table 3 reports the average true proportions and the proportions of the four categories estimated by Gradient Boosting and iSA. As expected by the random sampling, training and test sets have the same proportions of categories. The proportions of categories estimated by the two methods are all very similar to the true values. The Mean Absolute Error (MAE) shows that Gradient Boosting is slightly better than iSA.

8.1 Unbalanced Experiment

The setting of the second experiment is identical to the previous one, except for the sampling rate of the test set, that is, we sampled with unequal probability for each category. We do this with the aim of sampling very different test and training sets

Table 3 Results of the Balanced Experiment. True proportions (the ones used for training and the total ones) and estimated proportions of the four categories averaged over runs. Last column reports the Mean Absolute Error (MAE, i.e. lower is better).

	Info-Generic	**Info-Data**	**Action-Politics**	**Territ.-Action**	***MAE***
True Train	*0.724*	*0.104*	*0.097*	*0.075*	
True (Total)	0.723	0.105	0.097	0.075	
Grad. Boost.	0.759	0.092	0.083	0.066	*.075*
iSAX	0.681	0.107	0.132	0.078	*.157*

Table 4 Results of the Unbalanced Experiment. True proportions (the ones used for training and the total ones) and estimated proportions of the four categories. Last column reports the Mean Absolute Error (MAE, i.e. lower is better).

	Info-Generic	**Info-Data**	**Action-Politics**	**Territ.-Action**	***MAE***
True Train	*0.540*	*0.180*	*0.180*	*0.100*	
True (Total)	0.722	0.105	0.097	0.076	
Grad. Boost.	0.615	0.141	0.156	0.087	*.217*
iSAX	0.497	0.164	0.234	0.104	*.464*

with respect to categories' proportions.

The proportions of sampled documents for each topic together with the results are reported in Table 4. The Unbalanced Experiment reveals a big difference in performances between the two approaches. Both approaches fail to adequately estimate the true proportions of categories, but the Gradient Boosting approach performs better than iSA since estimated proportions are closer to the true ones.

9 Conclusions

Beck et al. (2013) identified cosmopolitan forms of action in (1) environmental associations (including green consumerism); (2) sustainable cities networks, focusing on sustainable territorial planning; and (3) innovation consortia (including "green" entrepreneurs and/or products). These "cosmopolitan communities of climate risk" seem quite underrepresented in our sample. Actions as such are addressed only in marginal ways. Through hand coding we found very few tweets that talked about concrete actions, potentially falling into the clicktivism/slacktivism (Morozov, 2009) phenomenon, i.e. many users retweeted but actually there is little concrete commitment (e.g. laboratory/workshop as territorial actions) or, in general, a willingness to act. For those who consider that sustainable development should be promoted by a global community as well as by technologically mediated commitment (Groves, 2019) - orienting individual and collective actions through shared practices

- it would have been desirable to find more of those practices in social media communication on 2030 Agenda. Nonetheless, in our data, actions are promoted in a limited number of tweets, an indication that some communities are able to connect global problems to local challenges, showing how the global impact of the climate crises is translated into specific territorial actions (Boyom et al., 2016). From this perspective, it there seems to be further untapped potential to promote exchange opportunities enabling different networks dialogue and facilitating reflexivity and mutual learning (Boström et al., 2017).

Overall, our Classification Supervised Model has produced satisfactory results, both for 2 and 4 categories, showing the importance of the variables deriving from textual analysis and those coming directly from Twitter. In this sense we want to stress the importance of data mashup, since our model can easily include variables of different type and source and, unlike iSA, we can observe the importance of each features. Furthermore, the comparison between the two methods (xgboost and iSA), related to the hand-coded subsample, showed that the Gradient Boosting achieves better results in terms of MAE.

A possible development of this research could include word embeddings in the features selection phase. In this way, features selection would focus on the semantic word similarities, rather than on the term frequencies. Addressing the research in this direction and, more in general, defining machine learning models with better prediction ability will produce more reliable analysis. Further research will address these issues.

Although we are aware of the limitations of this kind of studies (representativeness, noise, bots, etc.), we believe that probabilistic short-term models applied to the empirical observation of human behavior in large datasets mined from social media could extract social knowledge, representing a great opportunity for social scientists (Lauro et al., 2017).

References

1. Beck, U., Blok, A., Tyfield, D., Zhang, J.Y. (2013). Cosmopolitan Communities of Climate Risk: Conceptual and Empirical Suggestions for a New Research Agenda, *Global Networks*, *13*(1), 1-21.
2. Boström, M., Lidskog, R., & Uggla, Y. (2017). A reflexive look at reflexivity in environmental sociology. *Environmental Sociology*, *3*(1), 6-16.
3. Boyd, D. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679. doi:10.1080/1369118X.2012.678878.
4. Boyom, C., Callens, S., & Cherfi, S. (2016). Cultures of sustainability according to Ulrich Beck scheme: territorial strategies for electromobility. *Journal of Innovation Economics Management*, (1), 135-158.
5. Ceron, A., Curini, L., & Iacus, S. M. (2016). iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, *367*, 105-124.
6. Ceron, A., Curini, L., Iacus, S.M. (2017). *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge: New York.
7. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. (2019). *xgboost: Extreme Gradient Boosting*, R package version 0.81.0.1, 1-4.
8. Friedman, J.H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378.
9. Groves, C. (2019). Sustainability and the future: reflections on the ethical and political significance of sustainability. *Sustainability Science*, 14(4), 915-924.
10. Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229-247.
11. Lauro, N.C., Amaturo, E., Grassia, M.G., Aragona, B., Marino, M. (Eds.). (2017). *Data Science and Social Research: Epistemology, Methods, Technology and Applications*. Springer: Heidelberg.
12. McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological methods & research*, 46(3), 390-421.
13. Morozov, E. (2009). The Brave New World of Slacktivism. *Foreign Policy*. Available online: https://foreignpolicy.com/2009/05/19/the-brave-new-world-of-slacktivism, accessed 15/03/2019.

14. Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.
15. Sampson, J., Morstatter, F., Maciejewski, R., & Liu, H. (2015). Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In *Proc. of the 26th ACM Conf. on Hypertext & Social Media* (pp. 237-245). ACM.
16. Surian, A., & Sciandra, A. (2019). City Prosperity Index: a comparative analysis of Latin American and Mediterranean cities based on well-being and social inclusion features. *Book of Short Papers ASA Conference 2019, Statistics for Health and Well-being*, Padova: Cleup, pp. 210-214.
17. Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110, 298-310.
18. United Nations General Assembly. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*, Resolution 70/1.