A mixture of Topic Modeling and Network Analysis. The case-study of climate change on Twitter

Cristiano Felaco[1], Rocco Mazza[2], Anna Parola[3]

[1]University of Naples "Federico II" – cristiano.felaco@unina.it

[2]University of Naples "Federico II" – rocco.mazza@unina.it

[3]University of Naples "Federico II" – anna.parola@unina.it

Abstract

The paper proposes a semi-automatic labeling of topics extracted with a Topic Model using the tools of Social Network Analysis. The aim is to attach a label to every topic studying the terms-topics network structure. This method performs a semi-automatic topics labelling by using Latent Dirichlet Allocation model, integrating the network approach with topic generative model. LDA allows to extract latent topics and Social Network Analysis' tools permit to delineate the neighborhood of each topic, fostering a stronger interpretation of the meanings of the topics through the analysis of the extracted topics and documents' terms. To better show the joint use of Topic Model and Network Analysis, we present a case-study of how young people feel the climate change through the messages of user @Fridays4future extracted by International Fridays For Future Twitter account.

Keywords: Text Analysis; Topic Modeling; Social Network Analysis; Climate change; Young.

1. Introduction

The main aim of the paper is to show a semi-automatic labeling of topics using a mixture of Topic Modeling and Network Analysis. Specifically, our approach is builds upon different phases: (1) corpus preprocessing to build document-term matrix; (2) applying the model to extract topics and to identify the latent themes within the contents collected; (3) improving the interpretation of the meanings of the topics through the use of network analysis's tools in order to better identify the neighborhood of each topic, and, finally, (4) studying the semantic structure that links together the emerged themes with the intent to understand the semantic relationships between the extracted topics and documents' terms. The starting point is the Latent Dirichlet Allocation model. LDA is applied for the latent theme generation, starting from collected documents. At the base of the LDA we find this assumption: a) documents are represented as mixtures of topics, where a topic is a probability distribution over words, as a generative and Bayesian inferential model; b) the topics are partially hidden, latent precisely, within the structure of the document. LDA allows to infer the latent structure of topics through by recreating the documents in the corpus considering the relative weight of the topic in the document and the word in the topic, in an iterative way.

The method proposed here is adopted to flesh out understanding of the public perception of climate change and its impact on human life. We built the textual corpus through the International Fridays For Future Twitter account, by using messages of user @Fridays4future.

This paper is structured as follows: section 2 show the method and procedures used; sections 3 illustrates the case study of perceptions and stances of young people on climate change' impacts, and section 4 more specifically discusses the results of the research.

2. Method and procedures

The methodological procedure consists of a series of steps: 1. Tweets extraction and pretreatment; 2. Topic model application; 3. Textual network analysis. The analysed tweets were extracted from FridaysForFuture official twitter account by using *Application Programming Interface* (Shi, 2011), from now API. These applications allow to access to social networks data by specific statements (Charu et al. 2015; Puschmann, 2017). The contents were extracted indicating some features like time periods, content type (status, documents, image, link, etc.). We downloaded all user's tweeter status (only document type) and build our corpus. Starting from corpus we operated pre-treatment operations: 1. Text normalization; 2. Empty words (stopwords) and numbers removal; 3. Links removal. In the end we generated *document-term* matrix and we removed sparse words and empty documents. On this matrix we applied a model to detect latent topic from and reduce the information to plot on the network. In the case in which corpora gather a great amount of information it becomes difficult to trace back the semantic structure constituting it, especially when the aim is to identify the latent topics that constitute the textual collections under analysis. In order to efficiently perform this task it is possible to use specialized tools that solve the issue by way of offering a simple yet statistically robust solution.

2.1 LDA topic modeling

Below a brief illustration of model used. A topic model is a model that enables us to detect a series of topics within a collection of texts: these topics are drawn out by the distribution of words and how they help in shaping, characterizing and semantically define the topic itself. In the vast reference literature we rely on the following definition (e.g., Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001), «Topic models are probabilistic latent variable models of documents that exploit the correlations among the words and latent semantic themes» (Blei and Lafferty, 2007). We can condense and divide the ideas at the core of this definition in three fundamental steps (Blei, 2009):

1. To uncover the underlying topical patterns that permeate the collection;
2. To annotate the documents according to those topics;
3. To use annotations in order to better reorganize, summarize, and search the texts.

Among the great variety of probabilistic topic model used to analyze the content of documents and the meaning of words (Blei et al., 2003; Griffiths and Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001), one worthy of attention is the Latent Dirichlet Allocation (from now LDA). LDA is a generative and unsupervised model, these characteristics were important for our choice to use it. We needed a method allows us to generate new texts, this texts represent the latent variables identified by the models. Our methodological objective is to study the semantic structure of this textual latent variables. This is impossible to reach with established methods, like clustering methods. The main reference to this peculiar model is certainly the work of Blei et al. (2003); moreover, the model has been extensively studied in Griffiths and Steyvers (2004), Heinrich (2005), Blei and Lafferty (2009), Berry and Kogan (2010), Blei (2011) and others. What follows in this section is an overview of the LDA topic model, largely based on the original authors' work. The ultimate goal is to infer this latent structure of topics. The model carries out this task by means of recreating the documents in the corpus and by estimating the relative weight of the topic in the document and the word in the topic in an iterative way/fashion. The Latent Didrichlet Allocation is a Bayesian model and assumes that document and topic distributions can be described by a Dirichlet distribution (Blei et al., 2003). Being the conjugate distribution of the Multinomial, it is very convenient to use it as a priori: it follows that it is an effective tool concerning the problems of model inference (Neapolitan, 2003; Balakrishnan and Nevzorov, 2003). Arranging the process schematically: from a Dir (α) Dirichlet distribution we carry out a random sampling which represents the distribution of the topics of a particular document; this topic distribution is θ; from θ, we select a particular topic Z based on the distribution; then from another Dirichlet distribution Dir (β), we select a random sample representing the word distribution of the topic Z, this distribution is φ, from φ we choose the word w. In unsupervised topic models the number of topics to extract from the corpus is a parameter to be defined a priori. However, there are various methods and algorithms for evaluating the model with the optimal number of topics (Blei and Lafferty, 2009, Wallach et al., 2009; Buntine, 2009; Chang and Blei, 2009). To find the optimal topics number we could run the model a defined number of times by entering various parameters and choose the best performing one. Among the various models, to measure the performance of the model we could rely on held-out data: "Estimating the probability of held-out documents provides a clear, interpretable metric for evaluating the performance of topic-related models related to other topic-based models as well as other non-topic-based generative models"(Wallach et al., 2009). In this paper the best model was selected through the harmonic mean method (Griffiths et al., 2005; Zheng et al., 2006; Wallach, 2006, Wallach et al., 2009; Buntine, 2009). It is possible to formalize the method in this way (Griffiths and Steyvers, 2004, p. 5231):

In our case, the data are the words in the corpus, w, and the model is specified by the number of topics, K, so we wish to compute the likelihood p(w|K). The complication is that this requires summing over all possible assignments of words to topics z [i.e., p(w|K) =Rp(w|z,K)p(z)dz]. However, we can approximate p(w|K) by taking the harmonic mean of a set of values of p(w|z,K) when z is sampled from the posterior p(z|w,K) (Kass and Raftery, 1995).

2.2. Textual network analysis and topics naming

The topic modeling allowed us to extract latent topics from document-term matrix, for each topic we have several words assigned to its with measured probability. This topic represents the latent themes in the corpus, but the model output das not reveal the links between words and topics. We need an approach to bring out and visualize the semantic structure of the texts. The aim is identifying a method in order to more easily explore the content of the textual data collection. We propose a technique for plot and analyse with specific indices the topics and the associated words. Our strategy is centered around the idea that we can use the network analysis techniques to systematize, summarize and visualize the structure emerged to results of fitted model. We built a terms-topics adjacent matrix using the first 20 terms allocated for each topic. This is a two-mode and non-squared matrix, sizes "*n x m*" with n cases on rows, and m affiliations on columns. Main goal on two-mode matrices study is to identify the presence of linkages established between the actors (the rows) and the events (the columns) plotting this on a network. We plot a network from this matrix to characterize the structure that links the two dimensions. It is spatially displayed by means of a graph in which the node represents single lemma or single topic, while the lines represent ties which link them together. In this view, the neighborhood of each topic is shaped by those lemmas that are linked to a specific topic; moreover, topics are not necessarily isolated node, but can be put into contact by those lemmas share links with more topics. To this aim, indices associated with the centrality are used to identify important concepts, associations among concepts, themes and so on. The indicators used are degree centrality and eigenvector centrality, these are specifically calculated for two-mode networks (Faust, 1997). In a bipartite graph the degree centrality represents the measure of affiliation both for actors and events. For the first one is the number of events with which they are affiliated and for the second is the number of actors affiliated with it. The degree centrality is an important measure for affiliation network, it expresses how much the actors are important because of their level of activity or the number of contacts that they have (link to events), and events are important because of the size of their memberships (link to actors). Our research prospective is centered on textual analysis, the degree centrality indicates how central is the lemma compared to the number of topics it belongs: it is based on the number of word-crossing over topics. The measure adopted in this work is normalized by dividing by the maximum number of ties possible, which in a graph of n nodes is n-1.

Equation 1 a. Degree centrality for actor b. Degree centrality for event.

$$C_D^{MN}(n_i)=\sum_{i=1}^{g+h} x_{ik}^{NM}=x_{ii}^{N}=a_{i+¿¿} \qquad C_D^{MN}(m_k)=\sum_{k=1}^{g+h} x_{ik}^{NM}=x_{kk}^{M}=a_{+k}$$

A general definition for eigenvector centrality is that the centrality of an actor should be proportional to two elements: the strength of the actor's relations to other network members and the centrality of these other actors (Faust, 1997). To explain this indicator for bipartite graphs is important understand how the centrality indices for actors are related to the centrality indices for events. The centrality of an actor is proportional to the centralities of the events with which it is affiliated, and the centrality of an event is proportional to the centralities of its members (Faust, 1997). The equations below express the centrality measure in terms of individual actor and event centrality indices. This formal relation clearly includes the duality between actor and event centralities (Bonacich, 1991), and the λ represent the largest eigenvalue. In textual network approach this centrality measure indicates how much a lemma is linked to topics which are themselves central (and vice-versa), and then very well connected.

Equation 2 a. Eigenvector centrality for actor b. Eigenvector centrality for event

$$C_E^{N}(n_i)=\frac{1}{\lambda}\sum_{k=1}^{h} C_E^{M}(m_k)\,a_{ik} \qquad C_E^{M}(m_k)=\frac{1}{\lambda}\sum_{i=1}^{g} C_E^{N}(n_i)\,a_{ik}$$

This approach outcomes a network plotted and the centralities measures to describe it. We used these tools to name each topic and describe the semantic structure that characterize the corpus. The topics naming is the procedure allow to interpreter and to label the texts generated by the model. The naming modalities are several, but there are two principal approaches: automatic and manual labeling. The difference lies in significant cognitive load in interpretation and in subjectivity of researcher. In the first approach the topics reading is presented with manual post-hoc labelling for ease of interpretation in research publications (Wang and McCallum, 2006). The second one focuses on reproducibility and automation of process (Mei et al., 2007). The method proposed in this paper is halfway between the upper methods. The subjectivity of the researcher is oriented by rational criteria that guarantee the generalizability. These refer to indices calculated starting from the affiliation matrix and the network structure. Substantially, our methods to topics naming is based on the study of neighborhood of each topic consisting of two *criteria*: 1. The lemmas linked to topics; 2. The reading of centrality measures outlined above.

3. #Fridayforfuture: perspectives of climate change

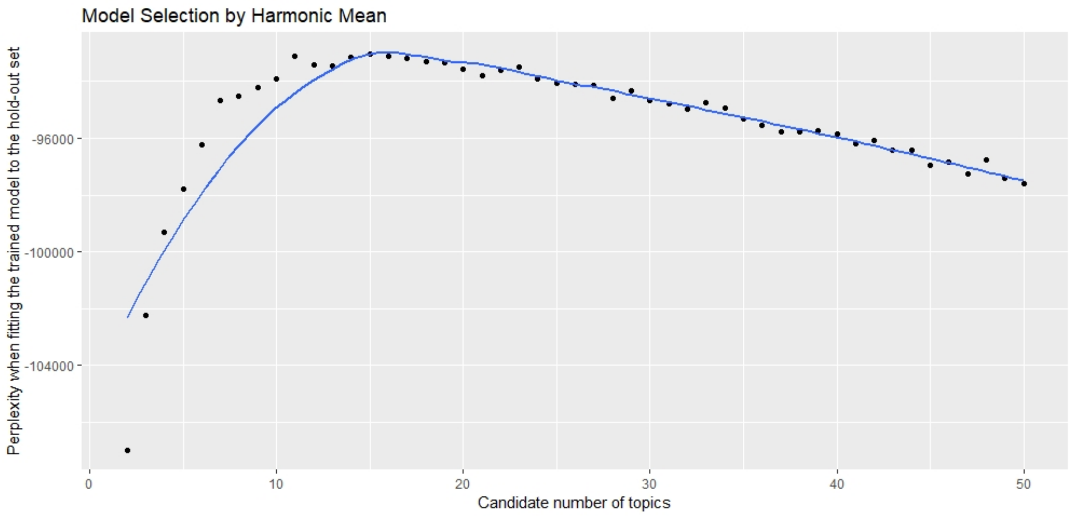
We present a case study conducted through the topic modeling and SNA in order to understand perception and stances of the climate change.

Climate change is inextricably embedded to individuals. In fact, in a cyclical perspective the individual provokes and suffers the climate change. Nevertheless, climate change still appears to be treated by many as a distant phenomenon, as well as temporally, socially, and geographically removed from our everyday experience (McDonalds, Yi Chai, Newell, 2015). The fact that it is not directly experienced (Morton, 2007, 2013; Boulton, 2016), it makes harder to empirical identify climate change, and then to accurately pinpoint the meanings perceived and the relevant themes about it. To this aim, the method proposed can be a suited strategy to better identify the main topics and then the meanings given to the climate change. Along with this, Twitter has been used to study climate change (Kirilenko and Stepchenkova, 2014; Jang and Hart, 2015; Cody, Reagan, Mitchell, Dodds and Danforth, 2015; Veltri and Atanaova, 2017; Segerberg and bennett, 2011; Kirilenko, Desell, Kim and Stepchenkova, 2017; Newman, 2017; Yeo et al.,2017). The media have been identified as an especially important agent in the formation of common-sense knowledge about climate change (Carvalho, 2010). In our case, even more actors prefer to turn to social media to disseminate information about climate change and mobilise support for action on climate change (Schäfer, 2012).

4. Results

The corpus was built with the statuses published by @Fridays4future, the official account of the international movement in the Twitter community. The textual dataset is composed of 3197 elements in English language, published between 25 of January 2019 and 11 of November 2019.

The figure 1 shows the calculation of the best model: the optimal number of topics is equal to the highest point of the curve, represented by points 15, drawn by the interpolation line.



lles

Fig. 1. Model selection.

The table below shows the list of topics extracted and the most relevant lemmas for each topic.

Topics are represented by a whole network of relations as a graph between each topic and lemmas (Figure 2). Specifically, the joint use of topic modeling and SNA: a) to enhance the understanding of the content of each topic through by analysing those lemmas linked to it; b) to analyze the relations between the topics through by the co-affiliation analysis, identifying those lemmas that are linked to more topics; c) to graphically represent these different kinds of relations.

Tab. 1. List of topics and words.

Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
today	day	join	gretathunberg	youth
new	one	want	fridayforfuture	alexandriav2005
striking	year	everyone	now	jeromefosterii
climatestrike	old	make	bangladesh	amp
students	Amagnussonn	parents4future	activists	lillyspickup
city	greta	climateemergency	right	now
Topic 6	*Topic 7*	*Topic 8*	*Topic 9*	*Topic 10*
strike	fridayforfuture	week	greenpeace	fridayforfuture
friday	gretathunberg	climatestrike	today	climatestrike
global	climatestrike	school	love	strikeclimate
see	today	schoolstrike4climate	ever	thank
next	jamiemargolin	strike	fridays	may
september	way	makichyana	luisamneubauer	nakabuyehildaf
Topic 11	*Topic 12*	*Topic 13*	*Topic 14*	*Topic 15*
climate	people	will	climate	world
change	young	future	just	time
crisis	support	can	strike4youth	action
actions	now	let	need	leaders
gretathunberg	climatechange	stand	climateaction	around
Countries	havenruthie	say	hollywildchild	must

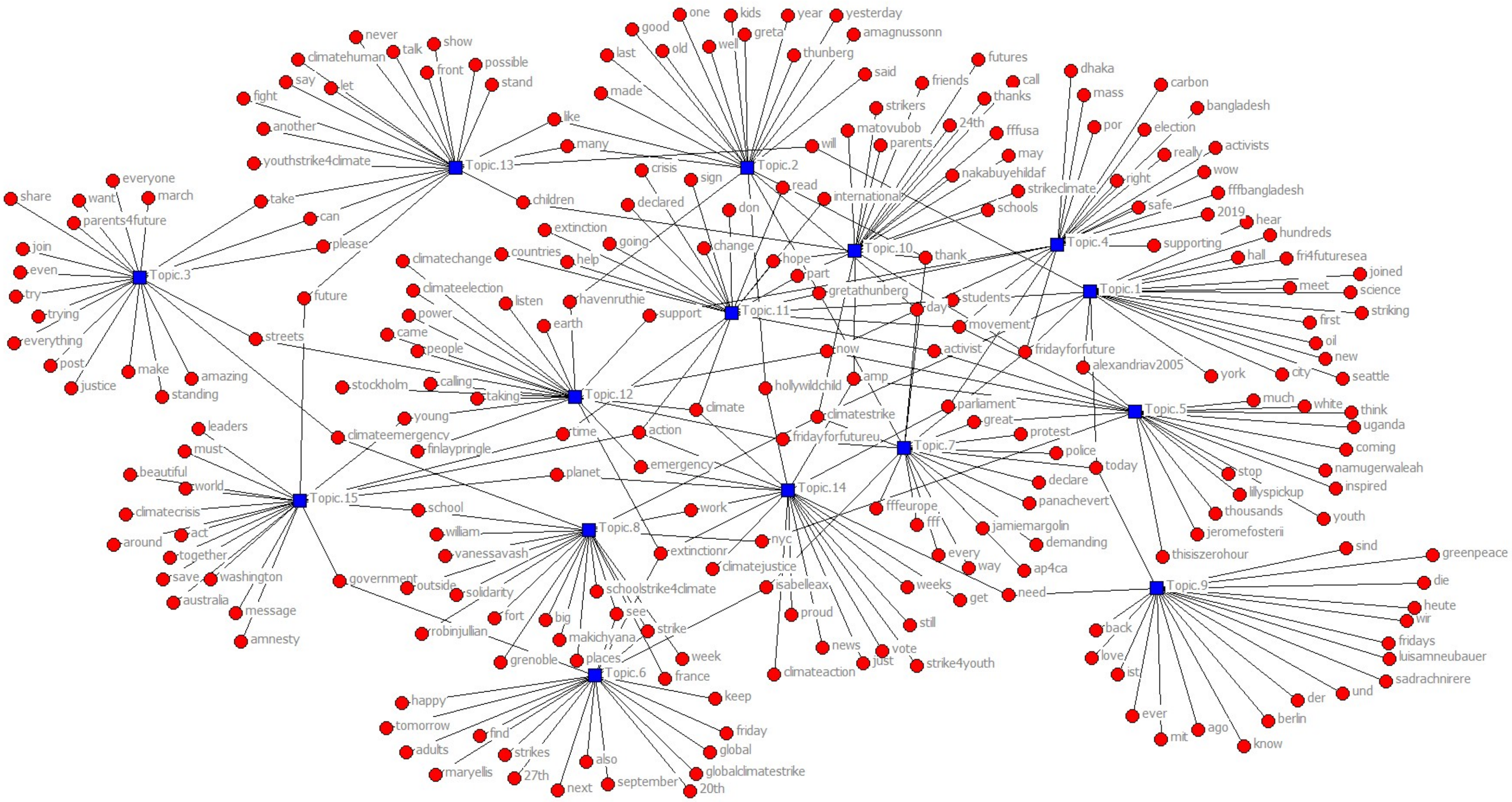


Fig. 2. Affiliation network of lemmas and topics (circles = lemma ; squares = topic)

Specifically, it emerges three main themes: a) climate change activists; b) strikes; c) future dimension. The majority of topics refers to events or movements linked to most popular *Friday for future* activists (i.e., Andreas Magnusson in topic 2, Greta Thunberg in topic 4, Haven Coleman (known as havenruthie) in topic 12, etc). Users approve the words and the actions of the activists in the battle to protect the environment and to defense social rights. Along with this, the strike is seen by young as a way to act and make their voices heard with the intent to influence or change their future, for a better life (topic 1). Other topics highlight, instead, strikers happened in the past (topics 6 and 7), underlining the past dimension. In addition, strike calls political solutions (topic 15) and collective actions (topic 3) that can make positive changes for the future (topic 13).

Beyond the analysis of the neighborhood of each topic, the use of indices associated with the centrality of lemmas can be useful to identify meaning paths and associations among different topics. To this aim, table 2 and 3 show those lemmas and topic with the highest values of normalized degree centrality and eigenvector centrality.

Tab. 2. Lemmas with higher values of Normalized degree centrality and Eigenvector centrality

Lemma	*Normalized degree centrality*	*Lemma*	*Eigenvector centrality*
climatestrike	0.267	fridayforfuture	0.239
fridayforfuture	0.267	gretathunberg	0.238
gretathunberg	0.267	climatestrike	0.213
today	0.200	thank	0.178
thank	0.200	movement	0.168
now	0.200	climate	0.165
movement	0.200	amp	0.162
amp	0.200	now	0.156
extinctionr	0.200	action	0.147
climate	0.200	today	0.135

Tab. 3. Topics with higher values of Eigenvector centrality

Topic	*Eigenvector centrality*
Topic.7	0.397
Topic.11	0.349
Topic.10	0.339
Topic.4	0.326
Topic.14	0.301
Topic.1	0.278
Topic.1	0.278

2	
Topic.5	0.270
Topic.2	0.264
Topic.8	0.181
Topic.15	0.175
Topic.13	0.133
Topic.6	0.131
Topic.9	0.084
Topic.3	0.074

Regarding the measures of centrality, we can observe that *fridayforfuture*, *gretathunberg*, *climatestrike* are the words most used in the tweets and those nodes according to their relative importance to other nodes. Topics with the highest values of eigenvector centrality are linked to those words with the highest values of the same measure: Topic 11 can be named as "Ideological actions for climate change" contains the main elements of the actions in defense of planet: movement, activist, change, extinction, crisis, climate, time and so on.

Greta Thunberg, the most representative environmental activist on climate change, plays a central role acting as bridge that allow the link between the topics 7 and 11. It indicates a continuity between the topic 11 that, as it has seen, reflects the ideal aspects of the movement Friday for futures, and the topic 7 that emphasizes the downside of the strike, the part more violent (protest, police, the case of Jamie Margolin). The latter topic may be named as "Practical actions for climate change".

A further central topic is the number 10 in which we find the figure of Nakabuye Hilda F., a Ugandan student striking for climate actions. This topic is both connected to topic 7 thanks those words that call directed and pragmatical actions (*climatestrike*, *fridayforfuture*), and to topic 11 through hope and part, that concern the feeling of belonging and the hope of a change. For this, topic 11 may be named as "Hope to make a change".

5. Conclusion

The paper proposed a semi-automatic labelling of topics extracted with a Topic Model using the Social Network Analysis tools. Specifically, this method performs a semi-automatic topics labelling by using LDA model integrating the network approach with topic generative model. This approach can intend to contribute a network visualization of topic modelling outputs.

The technique has been applied to the corpus of tweets extracted from FridaysForFuture official twitter account. The results show the topic of textual data collection. Through using jointly topic modeling and SNA it was possible to define the content of each topic and the relations between them. Specifically, the results show following topics: climate change activists, strikes and future dimension. Some lemmas are strategic in users' speeches: those with higher centrality values help us to better identify the connections between the topics and then to reconstruct the meaning's flows.

In conclusion, this analysis has demonstrated to be a valid technique to explore the content of the textual data collection.

Some limitations of the present study also need to be addressed. First, the tiny size of the corpus must be acknowledged. Future studies will consider the application of the technique to larger long corpus to allow also a greater impact of the technique and the collection of more information. Second, the next step could take into account the time dimension through by conducting a longitudinal study. Indeed, future research should investigate similar research questions using this mixture models refining the proposed technique.

References

Balakrishnan N., Nevzorov V. (2003). A Primer on Statistical Distributions, Wiley-Interscience.

Berry, M. W., Kogan, J (2010). Text mining: applications and theory, John Wiley & Sons.

Blei, D. M., Lafferty, J. D (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17-35.

Blei, D. M., Lafferty, J. D (2009). Topic models. In Text Mining (pp. 101-124), Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation, Journal of machine Learning research, 3(Jan), 993-1022.

Bolasco, S., De Mauro, T. (2013). L'analisi automatica dei testi: fare ricerca con il text mining, Carocci Editore.

Bonacich, P. (1991). Simultaneous group and individual centralities. Social networks, 13(2), 155-168.

Boulton, E. (2016). Climate change as a ‘hyperobject’: a critical review of Timothy Morton's reframing narrative. Wiley Interdisciplinary Reviews: Climate Change, 7(5), 772-785.

Buntine, W (2009). Estimating likelihoods for topic models. In: Asian Conference on Machine Learning (pp. 51-64). Springer, Berlin, Heidelberg.

Carvalho, A. 2010. Media(ted) discourses and climate change: A focus on political subjectivity and (dis)engagement. WIREs Climate Change, 1(2): 172–179.

Chang, J., Blei, D. (2009). Relational topic models for document networks. In: Artificial Intelligence and Statistics, pp. 81-88.

Charu R., Amit V., Bharath Y., Rama K., Kiran S. (2014). *API-FICATION*, Hcl Technologies, https://www.hcltech.com/sites/default/files/apis_for_dsi.pdf.

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. PloS one, 10(8).

Faust, K. (1997) Centrality in affiliation networks. In: Social networks, 19(2), 157-191.

Griffiths, T. L., Steyvers, M. (2002). A probabilistic approach to semantic representation. In: Proceedings of the annual meeting of the cognitive science society, Vol. 24, No. 24.

Griffiths, T. L., Steyvers, M. (2003). Prediction and semantic association. In: Advances in neural information processing systems, pp. 11-18.

Griffiths, T. L., Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.

Hallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D. (2009). Evaluation methods for topic models. In: Proceedings of the 26th annual international conference on machine learning (pp. 1105-1112). ACM.

Heinrich, G. (2005). Parameter estimation for text analysis. Technical report.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 289-296, Morgan Kaufmann Publishers Inc.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine learning, 42(1-2), 177-196.

Jang, S. M., & Hart, P. S. (2015). Polarized frames on “climate change” and “global warming” across countries and states: Evidence from Twitter big data. Global Environmental Change, 32, 11-17.

Kirilenko, A. P., & Stepchenkova, S. O. (2014). Public microblogging on climate change: One year of Twitter worldwide. Global environmental change, 26, 171-182.

Kirilenko, A. P., Desell, T., Kim, H., & Stepchenkova, S. (2017). Crowdsourcing analysis of Twitter data on climate change: Paid workers vs. volunteers. Sustainability, 9(11), 2019.

McDonald, R. I., Chai, H. Y., & Newell, B. R. (2015). Personal experience and the ‘psychological distance’of climate change: An integrative review. Journal of Environmental Psychology, 44, 109-118.

McDonald, R. I., Chai, H. Y., & Newell, B. R. (2015). Personal experience and the ‘psychological distance’of climate change: An integrative review. Journal of Environmental Psychology, 44, 109-118.

Mei, Q., Shen, X., and, Zhai C. (2007). Automatic labeling of multinomial topic models. InSIGKDD, pages490– 499

Morton, T. (2007). Ecology Without Nature: Rethinking Environmental Aesthetics (Cambridge: Harvard University Press).

Morton, T. (2013). Hyperobjects. Philosophy an Ecology after the End of the World (Minneapolis, London: University of Minnesota Press).

Neapolitan R. E. (2003). Learning Bayesian Networks, Prentice-Hall, 58 (4), pp. 1064-1082.

Nevzorov, V. B., Balakrishnan, N., Ahsanullah, M. (2003). Simple characterizations of Student's t2-distribution. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 395-400.

Newman, T. P. (2017). Tracking the release of IPCC AR5 on Twitter: Users, comments, and sources following the release of the Working Group I Summary for Policymakers. Public Understanding of Science, 26(7), 815-825.

Ponweiser, M. (2012) Latent Dirichlet allocation in R.

Puschmann, C., & Ausserhofer, J. (2017). Social Data APIs: Origin, Types, Issues.

Schäfer, M. S. (2012). Online communication on climate change and climate politics: a literature review. Wiley Interdisciplinary Reviews: Climate Change, 3(6), 527-543.

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. The Communication Review, 14(3), 197-215.

Shi, L., Zhong, H., Xie, T., & Li, M. (2011, March). An empirical study on evolution of API documentation. In International Conference on Fundamental Approaches to Software Engineering (pp. 416-431). Springer, Berlin, Heidelberg.

Steyvers M., Griffiths T. (2007). Probabilistic topic models. In: Latent Semantic Analysis: A Road to Meaning, eds. T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, Lawrence Erlbaum, page 427.

Steyvers, M., Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behaviour. Public Understanding of Science, 26(6), 721-737.

Yeo, S. K., Handlos, Z., Karambelas, A., Su, L. Y. F., Rose, K. M., Brossard, D., & Griffin, K. (2017). The influence of temperature on# ClimateChange and# GlobalWarming discourses on Twitter. Journal of Science Communication, 16(5), A01.

Wei, S. and Croft., W.B. (2006). LDA-based document models for ad-hoc retrieval. In SIGIR ’06, pages 178– 185.