Students' feedback on the digital ecosystem: a structural topic modeling approach

Annalina Sarra, Adelia Evangelista, Tonio Di Battista

1. Introduction

In March 2020, to contain the spread of the COVID-19 pandemic, almost all educational ecosystems (school, universities and private centres) around the world were forced to cancel face-to-face classes and replace them with didactic instruction online. Various and diversified methods of teaching delivered remotely were activated quickly. These solutions have undoubtedly had the purpose of ensuring the continuity of basic education and institutional activities, but they also made it possible to experiment, on large scale, didactic solution, mediated by screen, at design and didactic mediation level and interaction. The debate around the way educational systems reacted to the emergency is probably going to be a proper theme of investigation for next years. In this respect, (14), argue that the infrastructures for digital education that have been chosen to give a reply to the pandemic crisis, will redefine public education for the future. In addition, other scholars, see for example (2) and (6), have already carried out researches on screen-mediated didactics in the pandemic context. Their studies highlighted some essential specificities for a positive teaching-learning process, mainly related to the sociality and the possibility of working in cooperative environments, the possibility of co-building knowledge in an active way, within a community of practice. Following these lines of research, in this paper, we are aimed at capturing students' perspectives and perceptions on screen-mediated didactics during the pandemic emergency. Data have been collected through a survey, which consisted of open-ended questions administrated to students attending six teaching large courses, held by four professors in two different Italian universities (Macerata and Chieti-Pescara). In particular, in the research have been involved students who attended course of Educational Sciences degree (45 from the course of "Didactics" and 48 from the course of "Special Pedagogy"). The questionnaire was also administrated to students enrolled in the Primary Education degree programme: 230 from the course of "Technologies for Education and Learning", 230 from the course "Laboratory of Technologies" and 230 from the course of "General Education". Finally, there were students who attended the course "Didactics of Training", enrolled in the Pedagogical Sciences degree. All courses refer to the year 2019/2020. To circumvent the dilemma between the benefit of having open-ended questions and the cost associated with their analysis, we adopt, in this work, an unsupervised topic modelling approach. More in detail, we focus on Structural Topic Modeling (10), which is deemed a variant of Latent Dirichlet Allocation (1), suited to address the strict statistical assumption that all texts in the modelled corpus are generated by the same underlying process. The remainder of the paper is organized as follows. Section 2 describes the unsupervised topic modelling adopted, while Section 3 presents the results. Section 4 contains an interpretation of the main findings and the conclusions.

2. Methodology

Topic modelling, focusing on text mining and information retrieval, has received a lot of attention and gained widespread interest among researchers, in recent years, in many research

Annalina Sarra, University of Chieti-Pescara G. D'Annunzio, Italy, annalina.sarra@unich.it, 0000-0002-0974-0799
Adelia Evangelista, University of Chieti-Pescara G. D'Annunzio, Italy, adelia.evangelista@unich.it, 0000-0002-7596-9719
Tonio Di Battista, University of Chieti-Pescara G. D'Annunzio, Italy, tonio.dibattista@unich.it, 0000-0003-2139-7273

fields. The core idea behind topic models is that documents are mixture of multiple topics. One of the most used probabilistic topic modelling algorithm is the Latent Dirichlet Allocation (LDA) (1). In the LDA approach, documents are generated via 3-level hierarchical Bayesian structure, under which each document d_m is modelled a finite mixture over a set of K corpus-wide topics z_k (1) and each topic is modelled as a set of V words w_v. The generative process performed by LDA on a corpus of documents can be summarized as follows: for each topic z, choose the probabilities over words $\phi_z \sim Dir(\beta)$, where ϕ_z is drawn from a symmetric Dirichlet prior distribution with parameter β; for each document d, choose the probabilities over topics $\theta_d \sim Dir(\alpha)$, where θ_d is drawn from a symmetric Dirichlet prior distribution with parameter α; for each word w_{dn} in document d, choose a topic $z_{dn} \sim Multinomial(\theta_d)$ and choose a word $w_{dn} \sim Multinomial(\theta_{zdn})$. Being LDA a bag of words model, the order in which the words appear is disregarded. Additionally, although LDA is able to extract hidden topics from text document, it does not allow examining the relationship between document-level information and the content of a document model. This limitation can be overcome by using Structural Topic Modelling (STM), developed by (10). STM is a natural-language processing algorithm expressly designed to represent the effect of external variables on topical content (probabilities associated with words in each topic) and topical prevalence (proportion of different topics that occurs within documents). Through STM, it is possible to estimate a series of regression models that treat the prevalence of each identified topic as an outcome variable. The STM capability has been investigated in an extensive body of works, in the fields of economics, finance, political science, education, new media (see, among others, (12), (15)).

3. Results

The textual responses collected in this study were pre-processed using common steps for cleaning text data, including tokenization, lowercase conversions, stop-removal and lemmatizing/ stemming. Corpus preparation and cleaning were done using the *quanteda* package (4) in R (8). The final corpus contains 1354 documents. To avoid any possible inconsistences, we carried our topic analysis on the original texts, expressed in Italian language. The most frequently 20 words of the corpus are displayed in Figure 1. To extract hidden topics from the corpus, we used a STM package in R, developed by Roberts et al. (11). As argued by Roberts, for having semantically interpretable topics, words should tend to occur within response and their top keywords should be unlikely to overlap with keywords from other themes. The first analytical step was the identification of the appropriate number of topics. By triangulating different diagnostic measures (namely, held-out likelihood, residuals, semantic coherence and lower bound), 10-topic model was settled as the best option. In the topic labelling process, to come with topic labels that reflect the main themes in a clear and concise way, high probabilities (Highest Prob) words, frequency-exclusivity (FREX) words, Lift, Score metrics and top 10 representative words of each topic were used (Figure 2 and Figure 3).

The most interpretable Topics retrieved from STM were assigned to the following dimensions: "Physical space home" (Topic 1), "Lack of direct confrontation and relationship" (Topic 2), "Building the community: use of whatsapp", (Topic 3), "Ask question to the professor" (Topic 4), "Communication and learning tools" (Topic 5), "Feedback" (Topic 6), "Listen to the recorded lesson again" (Topic 7), "Interaction with teacher" (Topic 8) (see wordclouds displayed in Figure 4).

The top words occurring within Topic 1 (lessons, distance, face-to-face, value, added, home) stress how that topic is connected with a different reinterpretation of learning environment. In more detail, students underline two central aspects: the possibility to have more concentration at home but also some elements of distraction or linked to digital divide. Looking at the set of
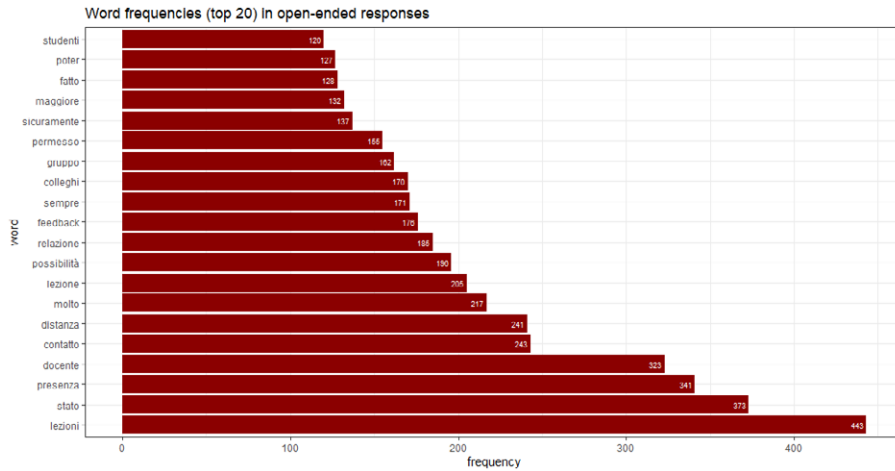
Figure 1: Word frequencies (Top 20) in open-ended responses

Topic 1	
Flex	valor, comodità, aggiunto, lezioni, spostamenti, dato, stato
Lift	costruzione, dovendo, fattore, persa, posto, recuperarla, riesco
Score	lezioni, valore, aggiunto, casa, seguir, comodità, spostamenti
Topic 2	
Flex	confronto, diretto, mancato, visivo, c'è, rapporto, altri
Lift	dispositivo, relazionarsi, visivo, confronto, diretto, adatto
Score	lezioni, valore, aggiunto, casa, seguir, comodità, spostamenti
Topic 3	
Flex	whatsapp, lavori, tenuti, gruppi, organizzazione, didattiche, videochiamate
Lift	accadeva, accordo, allungati, and, ansi, arricchendo
Score	whatsapp, tenuti, lavori, gruppo, gruppi, colleghi, organizzazione
Topic 4	
Flex	domande, dire, porre, me, disponibili, chat, secondo
Lift	accolto, adibito, all'esposizione, andata, apportato, apprendere
Score	domande, stato, maggiore, professore, dire, porre, chat
Topic 5	
Flex	supporto, canal, grazi, emotivo, svolto, telegram, app
Lift	aiutati, amicizie, elaborati, formativo, messaggi, organizzarci, poterci
Score	emotivo, telegram, canale, supporto, grazi, gruppo, tramite
Topic 6	
Flex	feedback, simile, disponibilità, minor, avvien, risponder, avviso
Lift	chiedeva, correzioni, eg, mail, ottima, preciso, riuscito
Score	feedback, simile, disponibilità, docente, stato, avviene, minor
Topic 7	
Flex	riascoltar, riveder, ascoltar, studentessa, registrazione, possibilità, registrare
Lift	affrontati, aggiunti, capita, dando, dell'ambiente, fondo, immediata
Score	riascoltar, riveder, registrare, registrazioni, seguir, possibilità, lezioni
Topic 8	
Flex	é, relazione, proprio, sì, stata, canali, diversa
Lift	abilità, abitudine, accaduta, adottata, affrontando, all'alta, all'aspetto
Score	relazione, docente, stabilir, dad, sì, diversa, cercato
Topic 9	
Flex	persona, dietro, crea, assolutamente, schermo, davanti, veder
Lift	rimaner, quindici, abita, accattivante, accogli, accontentare, accorgendo
Score	davanti, assolutamente, stare, schermo, persona, guardar, può
Topic 10	
Flex	compagni, contesto, nessuno, conoscenza, mancata, professori, pausa
Lift	adesso, compagni, contesto, correre, d'aiuto, erasmus, espormi
Score	mancata, compagni, professori, contesto, nessuno, erasmus, concetto

Figure 2: Top words for each topic according to highest probabilities, FLEX, LIFT and SCORE weighting
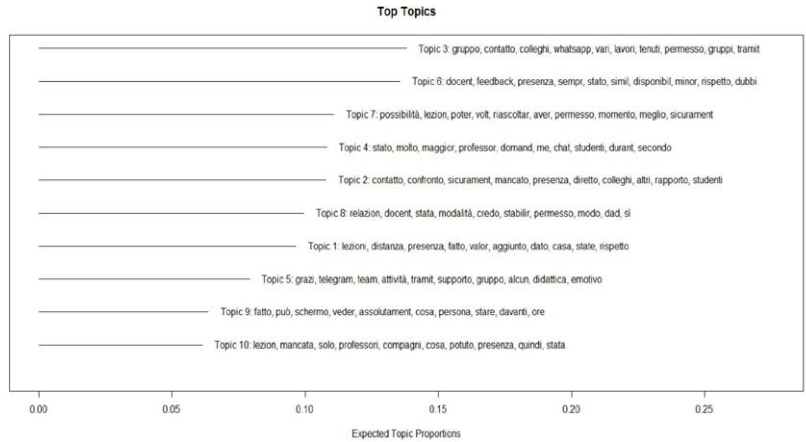
Figure 3: Top words associated with each topic resulting from structural topic modeling ($k = 10$)

Figure 4: Wordcloud: a) Topic 1: *"Physical space home"*; b) Topic 2: *"Lack of direct confrontation and relationship"*; c) Topic 3: *"Building the community: use of whatsapp"*; d) Topic 4: *"Communication and learning tools"*; e) Topic 6: *"Feedback"*; f) Topic 7: *"Listen to the recorded lesson again"*.

words linked to Topic 2 (contact, confrontation, absence, presence, direct), we are able to state that students think that interaction is somehow limited in the screen-mediated mode. Topic 3 focuses on the attempts made by the students of rebuilding the community or the contact with the other. Words associated to Topic 4 (questions, asking, available, greater, professor) recall the possibility for students to constantly ask questions to teacher. Terms immersed in Topic 5 refer to the online learning platform, perceived by students as essential for both supporting learning in an uncommon situation and as a space for discussion. Topic 6 captures the centrality of interaction and specifically of feedback and highlights how the teacher's feedback has not changed during the transition from face-to-face teaching methods to online mode. The top scoring words for Topic 7 clearly refer to the possibility of listening again to the lesson and of watching it more and more times, getting back to it in a recursive way. Finally the discussion in Topic 8, gives us the students' perception of having built a sound relationship with the professor. More challenging was to get insights from the last two dimensions characterized by less focused words. We also estimated the correlation between the identified topics. Except for "Interaction with teacher", the other topics are associated with at least a topic, meaning that they are likely to occur within the same documents. Finally, to complete the quantitative analysis of textual data, we incorporated the covariate information into topic modeling. Specifically, we estimated the topical prevalence by "teacher" covariate. The regression results support the causal impact of "teacher" variable that especially affects how Topic 2, Topics 5, 6 and 7 vary by document.

4. Discussion and Conclusions

The purpose of this study was investigating how students, who attended courses in two Italian universities, experienced online education during the coronavirus emergency. To this end, we used an unsupervised approach, based on the identification of latent topics, to automatically analysis open-ended questions. A throughout analysis of topic modelling results allows us to draw the following conclusions. By considering the perceptions in relation to blended environments, modellized by Chang and Fisher (3), we focus on the categories of "Interaction" and "Reply", which exploring to what extent communication is achieved from students' point of view and how students had felt about using web-based medium, respectively. Topics retrieved by the structural topic modelling analysis can be aggregated into three broad themes: perceptions related to the *physically of body and space*, perceptions related to *virtual relationships* and *communication and perception related to feedback*. Topic 1 and Topic 7 fall in the category "Spatiality and corporeity". In the distance learning mode, students recognized the undeniable advantages of being free from having to move: due to distance educational technologies implementation, remote learning is available to everyone, in any place. This aspect enables to stretch the same concept of access and participation and it has to be considered as an element of inclusion. Additionally, students reported the possibility of a greater interaction and participation during the lesson and the opportunity to listening again to the lesson and of watching it more and more times, getting back to it in a recursive way along time and in different moments. Under the umbrella of "virtual relationship" theme, there are Topics 2, 3 and 5. Based on the results of the topic modelling algorithm, we found out that students expressed that the filter of the screen was perceived as a barrier. In fact, even if online learning enables them to see each other and talk each other, it interrupted the relation flow that used to be experienced in a classroom. Finally, Topics 4 and 6 are the relevant themes for the broad category "Feedback". Throughout these topics, students underlined how the emergency remote education did not compromise the possibility of giving and receiving some feedback. Overall, the results of this study suggest the fluidity of contemporary education context: in other words, we are in front of a dynamic, hybrid educational context, with a weak structure, in continuous transformation (7). This feature,

exacerbated during crisis periods for the emergences of new obstacles and constraints, requires a rethinking of learning-teaching practises. A robust pedagogically and learning environment can be guaranteed by hybridizing the educational contexts. "Vertical blended", which provides for an alternation between moments of classroom teaching activity and remote teaching moments, must be accompanied by a "Horizontal blended", which integrates and hybridizes real and virtual, analogical and digital in a synchronous dimension (9).

References

[1] Blei, D. M., Ng, A. Y. , Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, pp. 993–1022.

[2] Carrillo, C., Flores, M.A. (2020). Covid-19 and teacher education: a literature review of online teaching and learning practices. *European Journal of Teacher Education*, **43**, pp. 466–487.

[3] Chang, V., Fisher, D. (2003). The validation and application of a new learning environment instrument for online learning in higher education, in *Technology-rich learning environments: A future perspective*, eds. M.S. Khine, D. Fisher , London: World Scientifics, pp. 1–20.

[4] Kenneth, B., Watanabe, K., Wang, H., Nulty, P., Obeng, A. , Müller, S., Matsuo, A. (2018) quanteda: an R package for the quantitative analysis of textual data. *Journal of Open Source Software*, **3** (30), pp. 774.

[5] Kuhn, K. D. (2018) Topics and trends in aviation incident reports. *Transportation Research Part C: Emerging Technologies*, **87**, pp. 105–122.

[6] Pereira, S.P., Fernandes, R.L., Flores, M.A. (2021) Teacher education during the covid-19 lockdown: insights from a formative intervention approach involving online feedback. *Education Sciences*, **11** (400), pp. 1–14.

[7] Pennycook, A. (2005). Teaching with the flow: fixity and fluidity in education. *Asia Pacific Journal of Education*, **25**(1), pp. 29–43.

[8] R Core Team (2022). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

[9] Rivoltella, P. C. (2021). *Apprendere a distanza. Teorie e metodi*, Raffaello Cortina Editore. Milano, Italia.

[10] Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M. (2013). The structural topic model and applied social science, in *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*.

[11] Roberts, M. E., Stewart, B. M., Tingley, D. (2019). stm: an R package for structural topic models. *Journal of Statistical Software*, **91**(2). https://doi.org/10.18637/jss.v091.i02

[12] Rodriguez, M. Y., Storer, H. (2019). stm: a computational social science perspective on qualitative data exploration: using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services*, **38**(1), pp. 54–86.

[13] Rothschild, J. E., Howat, A. J., Shafranek, R. M. , Busby, E. C. (2019). Pigeonholing partisans: stereotypes of party supporters and partisan polarization. *Political Behavior*, **41**(2), pp. 423–443.

[14] Selwyn, N., Jandrić, P. (2020). Postdigital living in the age of covid-19: unsettling what we see as possible. *Postdigital Science and Education*, **2**(3), pp. 989–1005.

[15] Zafari, B., Ekin, T. (2019). Topic modelling for medical prescription fraud and abuse detection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**(3), pp. 751–769.