Research and Teaching Public Communication of Science and Technology on Digital Data

Emanuele Di Buccio[1,2,3], Federico Neresini[3]

[1]Department of Information Engineering, Univesity of Padova
[2]Department of Statistical Sciences, University of Padova
[3]Department of Philosophy, Sociology, Education and Applied Psychology, Univesity of Padova

Abstract

In recent decades, there has been a growing interest among Social Science researchers in computational approaches; Computational Social Science and Digital Sociology are examples of these research directions. An interdisciplinary research field that can be framed within Social Science is Public Communication of Science and Technology (PCST), which examines how science and technology can affect contemporary society and how society can affect science and technology. The digitization of traditional media and the proliferation of other information channels, such as Social Media, provide new opportunities for PCST. This paper discusses the issues that need to be addressed to support PCST scholars, possible solutions to address them, and the integration of these solutions into a single platform that is being used to support research and teaching. Concerning teaching, the paper presents an example of how the platform can be used in the context of a university course.

Keywords

Digital Social Science, Research Platform, Public Communication of Science and Technology

1. Introduction

Social Science is a broad field of research that includes multiple disciplines, such as Sociology and Political Science. In recent decades, there has been a growing interest in adopting computational approaches to deal with the increasing amount of digitized content available [1, 2]. There is a large body of work that involves the adoption of Machine Learning (ML), and more generally AI-based or AI-inspired methodologies to support the research tasks of social scientists [2], to propose new interdisciplinary methodologies, or to rethink and possibly automate previous theories and approaches from a new perspective [3, 4].

This work will focus on an interdisciplinary field called Public Communication of Science and Technology (PCST) [5]. This field includes the "practice to make specialized knowledge available for the public" [6]; science communication "is used to inform, engage, persuade, change behaviors, and support better decision making [. . .] aims to lift the social, environmental and economic standing of a nation's people [. . .] It may also support the participation of citizens in setting the agenda for scientific research." [7]. Communication plays a crucial role in scientific

IRCDL 2024: 20th conference on Information and Research science Connecting to Digital and Library science formerly the Italian Research Conference on Digital Libraries, Bressanone, Brixen, Italy - 22-23 February 2024

emanuele.dibuccio@unipd.it (E. Di Buccio); federico.neresini@unipd.it (F. Neresini)

0000-0002-6506-617X (E. Di Buccio); 0000-0003-3918-2588 (F. Neresini)

research, e.g., to attract attention and to enhance its legitimacy in the eyes of its stakeholders and potential supporters [8, 9]. The public debate on science-related issues may begin when the debate among experts (scientists) is still "ongoing" – see, for example, the case of *cloning*. Even when there is a consensus among scientists, scientific issues become controversial when they become the subject of public debate. The mass media – both "old" media, such as newspapers, and "new" media, such as social networks – play an essential role in the public perception of scientific and technological issues and innovations. For example, they constantly propose different (interpretative) *frames* that may influence how the public perceives a particular issue. If, therefore, PCST activities take place mainly in the media arena, this means that they are characterized by the dual role of the mass media: on the one hand, they constitute a privileged space within which relationships between science and society occur; on the other hand, the media themselves contribute to fueling and shaping these relationships. For this reason, the analysis of media communication on science and technology represents an excellent opportunity for the social sciences.

This paper will consider research and teaching on PCST. We will reconsider the analysis on user needs conducted in the context of Digital Humanities [10, 11] and Social Science Research [3] and report on issues that need to be addressed to support these activities, possible directions to address them, and our past and current efforts in pursuing these directions. To show how we can adopt some of the proposed solutions for teaching, we will discuss a recent activity we carried out during a course for the Master's Degree in Communication Strategies.

2. Issues in PCST using Digital Data

Public Communication of Science and Technology encompasses several activities. In this paper, we will focus on media monitoring, which aims to follow the discourse – in the case of PCST, the discourse on science and technology – on one or more media channels. One of the channels traditionally followed is newspapers — in this sense, we can think of newspapers as "old" media compared to more recent channels such as social media platforms. Data digitization allows PCST scholars to test their research hypotheses on both "new" articles and historical newspaper data. Several research projects have focused on designing and developing digital libraries to preserve and provide access to historical newspaper archives; recent projects include NewsEye [12] and Impresso [13]. Research in PCST through newspapers is still relevant today, for instance, because it offers the possibility of investigating research hypotheses that require longitudinal studies. For example, suppose we wanted to follow the discourse of computing technologies or AI over several decades, starting in the 1960s. Social media are too recent to be a data source for such research questions. However, social media is an invaluable and necessary source for other research questions, such as studying the discourse on COVID-19 or Generative AI, newer viewpoints, or interactions specific to new platforms. Other media channels, such as vlogs or podcasts, are now available and can be used as data sources.

Therefore, PCST can primarily benefit from working with digitized data and, as we will discuss later, using computational approaches to uncover how the media present and "frame" issues, such as those related to science and technology. To achieve this goal, however, several issues must be addressed and are discussed in the remainder of this section.

2.1. Datasets and continuous media monitoring

Content analysis is a well-established practice in social science research. However, most previous studies have been based on samples, e.g. a subset of articles on a particular topic. While this is acceptable for some research questions, other questions require analysis of or comparison with the "entire population." For example, when studying the presence of articles on science and technology over time, working only with articles relevant to science and technology and looking at absolute frequencies would have led to an incorrect conclusion: that media pay growing attention to science and technology. However, it is not the case because the relative frequency is almost constant over time in the last decade [14]. Another reason for not working with samples is the definition of the object of interest. If we are interested in following the discourse of science and technology in newspapers, why not focus only on the "Science" and "Technology" sections? The reason is that we will miss part of the discourse: what about the debate on these issues within articles mainly focused on sport or business?

In the case of newspapers, studies are rarely conducted on all the newspapers available online: a subset of them is selected to include those that are representative of different spins. Even when a subset of newspapers (and sources in general) is selected, the size of the data, which is longitudinal in nature, may require a scalable platform to handle it. Media monitoring and approaches dedicated to handling news content are not new to IAR and NLP: the topic has been the focus of evaluation campaigns, workshops,[1] and projects; dedicated datasets[2] have been created. Therefore, PCST can benefit from these methodologies and resources.

Several platforms allow experts in other disciplines with limited programming skills to work with data; examples include Knime[3], Orange Data Mining,[4], and CorText [15]. Knime and Orange allow workflows to be implemented via "visual programming", specifically by connecting blocks — *nodes* in Knime and *widgets* in Orange; different types of blocks exist, e.g., for data collection, preprocessing, and analysis. CorText allows workflows to be implemented using several existing functionalities for data collection, subsetting, preprocessing, and analysis. While continuous data monitoring can be implemented using more advanced or custom functionality in Knime or by augmenting CorText with other libraries, these solutions may not be easily implemented by PCST scholars, that can benefit from an integrated environment, complex search functionalities, and (potentially very large) dataset export capabilities.

A different approach is taken by systems such as NOAM [16], The European Media Monitor [17], NewsEye, or Impresso. These systems aim to provide an integrated and ready-to-use environment. However, none of them meet all the requirements when it comes to continuous and ongoing monitoring (and not just archives), IAR, and support for the strategies PCST scholars use to conduct research (see section 2.3) or within teaching activities (see section 3.2).

2.2. Heterogeneous sources access and processing

A common methodology in PCST, given a research hypothesis, is to conduct a comparative study between different sources/channels. For example, we can study the same phenomenon,

[1] See, for example, https://research.signal-ai.com/newsir18/ and https://research.idi.ntnu.no/NewsTech/INRA/

[2] See, for example, https://catalog.ldc.upenn.edu/LDC2008T19 and https://trec-core.github.io/2018/

[3] https://www.knime.com/

[4] https://orangedatamining.com/

e.g. the public debate on AI, in different countries and use media channels such as newspapers and social media as proxies for public opinion. This requires the construction of corpora that are aligned in terms of the temporal dimension in different languages, from different channels, and possibly in different modalities. Podcasts, for example, can be a relevant source to study nowadays since there are many of them focused on news, and besides the content presented in the episodes, one could also look at the way the content is delivered, e.g., as done in [18], by considering vocal and conversational properties when predicting seriousness and energy.

Even when libraries are available to collect and preprocess different types of data, such as those mentioned above, access to the channels can be a problem. Newspapers such as The Guardian[5] or The New York Times[6] provide Web APIs. However, services for collecting data from (some) social media that used to be freely available for research purposes are no longer available or, if available, require substantial fees to download large amounts of data. This can severely limit research on these channels.

In addition to access to sources, another aspect to consider is content heterogeneity. Even when considering the same modality, e.g., text, different channels may require different methods. A well-known example is the case of topic extraction methods, such as Topic Modeling (TM) algorithms, whose effectiveness may be affected by document length [19], which may vary in datasets consisting of microblog or forum posts such as Reddit, or when different sources are considered simultaneously [20].

2.3. Workflow support

When dealing with experts in other fields, such as Humanities or Social Sciences, one central issue is supporting their *workflows*. Those experts alternate quantitative and qualitative approaches to investigate their research questions or, more in general, to accomplish a task.

A discussion on this aspect is reported in [11]. Even if the contribution is in the context of the NewsEye project and about the study of historical newspapers, the authors provided an abstraction of the problem and proposed a workflow that can adopted as a conceptual tool. The authors discuss how an interdisciplinary digital hermeneutics workflow is necessary to pursue the direction of interdisciplinary research that does not consider only the distinct points of view – the one by humanists and the one by computer scientists – but a joined view. The goal is to move us away from "supporting *their* workflow" and towards an interdisciplinary approach. For instance, it is not always possible to frame a task in "simpler" subtasks and then later translate them in pipelines: this is not how humanists (and in our case PCST scholars) proceed. For this reason, [11] proposed a workflow that considers three main aspects: (a) data, (b) iterative qualitative analytical steps over the data, and (c) critical reflection on data, algorithms, and tools.

As for (a), besides the need to focus on specific subcorpora, another critical aspect is the curation of the data. In the event of historical newspapers as in [11], tasks related to this point included extracting content and metadata from scans or images via OCR technologies. Based on our experience with the more "recent" online newspapers, we can add that automatic techniques robust to diverse templates and structures of the pages are required when scaling on the number

[5] https://open-platform.theguardian.com

[6] https://developer.nytimes.com

of sources. Another point stressed in [11] is the importance of metadata to get the context of the themes under investigation, which is crucial for deriving meaning from the data.

As for (b), a key aspect is the iterative approach and the high level of interaction required with the data. In this case, search can be a specific tactic within a more complex strategy to accomplish a task. In [10], the authors discuss some relevant tasks when working with historical newspapers and useful digital interface functionalities to accomplish such tasks:

1. filtering and searching by full-text queries and metadata such as time or newspaper;
2. identification and disambiguation of named entities;
3. identify the first occurrence of words or expressions;
4. study the change in meaning of words over time;
5. extraction of themes (topic in TM), interaction with theme descriptions (labels) for their interpretation and possible refinement, access to a representation of a document based on themes, visualization of the prominence of themes over time;
6. advanced search features such as relying on Boolean or regular expressions.

In addition, [11] mentioned the importance of improving search beyond keywords since some concepts are complex to express by a set of keywords, even if the suggestion/extraction of new relevant keywords – such as named entities extracted from the subcorpora – could help.

The last point (c) is crucial for computer scientists because it requires “openness and transparency of methods and tools” [11]; this is important both for reproducibility and to make explicit the assumptions underlying methods and algorithms and the role these assumptions play in investigating the experts’ research questions.

We observed analogous needs when interacting with PCST scientists [21, 22]; they require:

- better support for IAR, going beyond keyword-based search;
- ways to easily incorporate new and possibly heterogeneous sources for new perspectives on the public perception of science and technology issues;
- to switch from one (quantitative) strategy to another, to return to a more qualitative analysis, and then to perform further interactions;
- to compute a set of consolidated or new indicators, e.g., the “risk indicator” [23], on the subcorpora identified after several iterations.

Digital platforms to support research and teaching in these areas should be able to provide these functionalities in an integrated environment or at least facilitate the “implementation” of complex workflows that are not necessarily linear, but that can support different strategies and the alternation between quantitative and qualitative approaches to data analysis.

3. Towards Supporting Research and Teaching in PCST

3.1. Research

In Section 2, we identified three main issues: (i) datasets and continuous monitoring; (ii) heterogeneous sources access and processing; (iii) lack of supporting workflows for PCST scholars. In this section, we will describe how we are currently addressing them in an interdisciplinary project called TIPS[7] (Technoscientific Issues in the Public Sphere).

[7] https://www.tipsproject.eu/tips/

The first issue, i.e. not to limit the research to samples and to carry out continuous monitoring, has been addressed by designing and developing a modular software platform, presented in [21, 22], that stores and provides access to articles collected from fifteen newspapers in different languages; for three of them, archives are available that go back to the 1980s, obtained by complementing test collections, such as The New York Times Annotated Corpus, or collected through the available Web API, as for The Guardian. All newspapers are still being monitored so that we can work on more recent issues. Continuous monitoring required engineering effort to design and implement a robust architecture. We have integrated search, PoS tagging, named entity extraction, and topic modeling into a single platform that meets all of the requirements discussed in the section 2.1 for continuous and ongoing monitoring, IAR, and support for the strategies used by PCST researchers to conduct their research. Access to the platform requires authentication; credentials may be requested for research purposes. To support reproducibility, users can download metadata of the documents used for their analysis; the metadata includes the article's URL, which allows access to the full content via the original source.

Regarding the second issue, i.e. heterogeneous source access and processing, we designed the platform to work with arbitrary documents and different languages. The platform already stores and provides access to aligned corpora in different languages, which allowed us to perform comparative studies between different countries. Even if we do not monitor social media platforms, existing datasets can be easily included and processed by the existing pipeline. How to replace channels like the Twitter API is an open question and a solution has not been found yet. As for the robustness of the algorithms when working with heterogeneous data, such as [20], we have built temporally aligned test collections for different topics — e.g., DNA and AI — from 2010 to 2022 using data collected from social media and the news; they will be used to conduct experimental evaluations. The involvement of the PCST scholars provides a unique opportunity to gain insights into the effectiveness of these approaches on "real" tasks through qualitative evaluation.

The third issue, i.e., supporting workflows for PCST scholars, is the central aspect we are working on. The current functionalities already support filtering and search, named entity extraction, identifying the first occurrence of words and expressions, and advanced search features like regex or boolean constraints. A dedicated functionality in the platform allows the extraction of the top-named entities and nouns for a given query, thus helping users express their needs better. Additionally, users can work on specific subcorpora, obtained by boolean queries, directly within the platform, thus utilizing all the functionalities and indicators available without relying on other platforms or libraries. Some workflows are already supported, e.g.,

```
search → identify a subcorpus
→ topic extraction on the subcorpus
→ topic description and top docs per topic analysis
→ analysis of the evolution of topics over time
→ identify a subcorpus using a subset of the topics and re-extract them
```

Even if the platform supports multiple iterations, the above workflow is limited to search and TM. More articulated workflows can be integrated into the platform, e.g., that proposed in [24] to study the case of energy transition in Italian newspapers; that workflow involves additional techniques such as named entity recognition to identify prominent actors, and graph-based representations obtained from the articles' content to study actors relationships. Section 3.2

will present another possible workflow for PCST scholars and students.

Dealing with large amounts of data in terms of information access and automatic extraction of valuable and usable representations is not the only reason to introduce experts in PCST to computational approaches. Another reason is the increasing attention some of these Computer Science disciplines are receiving in the public sphere today. AI is becoming a prominent topic in the media and for political institutions, thanks to the progress achieved through new (computer) architectures and models and their widespread potential applications. Emerging technologies, such as AI-based chat-bots, are becoming controversial for their potential future impact on society, and institutional organizations have proposed specific regulations.[8] Sociologists and communication experts may be directly affected by these emerging technologies, and their role may be critical in discussing or communicating the implications for society. For example, when considering Generative AI, one concern is whether automatically generated content will become dominant. How might that affect society? In addition to the impact on political orientations or public perception, could other aspects be affected? These types of questions are relevant to PCST research activities that focus on the impact of digital technology on society and social interactions. These questions can also be the subject of teaching activities for future communication and sociology professionals, as we will discuss in the next section.

3.2. Teaching

The last remark on section 3.1 allows us to introduce another aspect, i.e. the introduction of students of social and communication sciences to topics in IAR, NLP and ML. This need has been the rationale for interdisciplinary courses such as those in Digital Humanties and Computational Social Science, which have been offered for several years. As discussed in [11] in the context of research on historical newspapers, "historians need to acquire new skills, especially in the practice of (digital) hermeneutics, which refers to the interpretation and understanding of large, digitized or digitally born data sources." The same is true for students (and scholars) in PCST. To promote an appropriate level of understanding of IAR, NLP, and ML topics, and to foster debate among the "future" communication professionals, it is imperative to introduce students to these concepts and also to some of the current and possible future implications. Following this direction, in this section, we will present a teaching activity we carried out on text classification to identify articles on emerging technologies; moreover, we will discuss other activities we plan to carry out to complement the first one.

As an example of teaching activities that might benefit students in Sociology and Communication, we report on recent experience in the course of Digital Sociology in the Master Degree of Communication Strategies at the University of Padova. The objective of the course is to introduce the students to epistemological and methodological issues concerning Digital Social Research, to digitalizing traditional methods (e.g., web surveys), data-driven social research, and making social research on Social Media and through Social Media, and digitalized newspapers.

As part of the course, the students were presented with how to represent unstructured data, such as newspaper articles, and ML techniques to analyze them, more specifically, supervised text classification on a specific object of study: emerging technologies. The task was framed

[8] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

as a binary classification problem where the goal was to determine whether a document was about emerging technologies. In the first lecture on this activity, the students were introduced to the *demarcation problem*, i.e., to the problem of determining the object of interest for the study; the goal was to identify a set of criteria then used to classify a document concerning relevance to emerging technologies. Even if the lecturers determined some criteria before the beginning of the activity, those criteria were not presented to the students, who were left to come up with the criteria and agree on those that must be adopted. The result of this activity was the following criteria: (i) a technology framed as "new" is mentioned; (ii) some impacts, i.e., changes, on society are described; (iii) a technology related to scientific research is mentioned.

After determining the criteria, we built a dataset using articles from six English newspapers[9] published from January 1, 2016, to November 7, 2023. The articles were retrieved by merging the results from the following queries:

- "emerging technology" OR "emerging technologies"
- chatgpt
- "fusion energy"
- "genome editing"
- "neuralink"
- "self driving car" OR "autonomous car" OR "driverless car" OR "robotic car" OR "google car" (and the corresponding version with "cars")

The expression between double quotes is interpreted as phrase queries (the constituting words must occur near each other as in the string). We considered candidate non-relevant documents those not returned as results for those queries. We set as an additional filter that articles must be relevant to Science and Technology issues, according to a previously developed classifier — see [21] for details on the manually labeled dataset and [22] for the approach used.

Then, we extracted a random sample of 658 documents that covered the diverse queries and (possibly) non-relevant documents; the sample was explicitly created, maintaining a balance between the two classes. The sample was then delivered to the students, who manually labeled 14 documents, each using the criteria. A total of 46 students were involved in the activity. Along with labeling relevant and non-relevant documents, students were asked to perform a quality check to identify duplicates or documents with incomplete text due to our extraction procedure. After labeling the assigned 14 documents, students were asked to label the documents from another student. After the labeling was performed, the students were divided into groups. They were asked to discuss the given labels and to agree on each document label, paying particular attention to problematic cases. The results were then provided to the lecturer. Some documents were removed because of the quality check; the resulting dataset consists of 642 documents.

In the following lecture, the students were introduced to fundamental notions on Computer Science, such as the notion of “algorithm”, and on ML learning, focusing on supervised text classification. The experience of labeling was instrumental in this subsequent lecture. Part of the lecture was devoted to presenting the effectiveness of some classifiers trained on the labeled dataset produced by the students. Even if the resulting dataset was small, we trained several classifiers using 5-fold cross-validation as a proof of concept. We used the JSAT Library [25],

[9] Mirror, The Guardian, The Telegraph, The New York Times, The Times of London, and The Financial Times

Table 1
Effectiveness of Text Classifiers for identifying articles on Emerging Technologies

Classifier	AUC	F1	Precision	Recall
MNB	0.837	0.771	0.832	0.721
LRDCD	0.809	0.766	0.770	0.764
Stacking	0.829	0.770	0.808	0.738

the one currently adopted "in production" in TIPS. The tested classifiers were Multinomial Naive Bayes (MNB), Logistic Regression with Coordinate Descent Methods (LRDCD) [26], and Stacking [27] of these two classifiers; the focus on these approaches was motivated by the previous encouraging results observed for classifying Science and Technology articles [22].

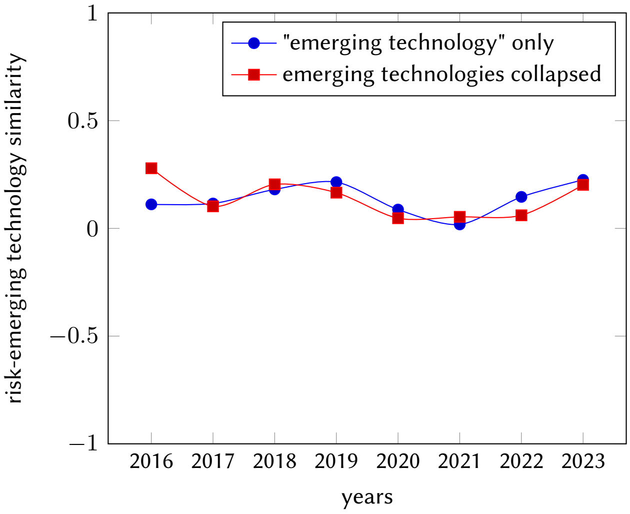
The subsequent lecture was devoted to current and possible future implications of ML approaches and technologies, e.g., approaches and technologies using behavioral data and Large Language Models, on society. The participation of the students and the constructive interactions were perceived as an indicator of a positive and valuable experience, and the main methodological aspects related to IAR, NLP, and ML were acquired by the students.

The students were then given a new sample to label and, later, the predictions based on the most effective classifier (the one using Stacking) to check the predictions' correctness. The labeling procedure followed the same approach adopted in the previous phase: labeling 14 documents and then discussing the labeling within the group. In the second phase, we got fewer labeled documents (383). As for comprehensiveness, Table 1 reports the results of the three classifiers on the full dataset; article URLs and labels of the adopted dataset are available in [28]. The obtained results suggest additional work should be done, e.g., increasing the size of the labeled set, before using the classifier for research purposes.

Text classification is one of many topics that can benefit teaching activities. The availability of classified and indexed longitudinal corpora allows us to present methodological aspects, such as the importance of working with the entire population to monitor some phenomena — see the example of the erroneous interpretation related to absolute frequency reported in Section 2. Another activity might be to show some representations obtained from the classified data. For example, we might ask if the perception of risk when discussing emerging technologies has changed over time. Can NLP and ML help? We might extract temporal word embedding representations and compare the distance among the embedding representation of the terms "risk" and "emerging technology" over time.

As a proof of concept and a basis for a possible activity to carry out for the next edition of the course, we considered all the documents answering the queries reported above and merged the obtained results; 80494 documents constitute the resulting subcorpus. We preprocessed them, replacing all the terms in the query constituted by more than one token, e.g., emerging technology, with a string to denote the entire expression; we then transform the text using lemmas instead of the original words — we used Stanza [29] for the extraction of the lemmas. Then, we trained a model using 100 dimensions to represent each word, 5 static iterations, and 5 dynamic iterations as suggested in [30]. We then computed the distance (cosine similarity) between "risk" and "emerging technology" over time; the trend is reported in Figure 1, specifically

Figure 1: Similarity between the term “risk” and “emerging technology” over time; *emerging technology collapsed* refers to the case where all the terms reported in the queries were replaced with “emerging technology” before the training.



the line connecting points depicted by circles. The other line (points depicted as squares) refers to the similarity between the term “’risk” and the term “emerging technology”, when in the pre-processing step, all the terms used in the query – “chatgpt”, “fusion energy”, “genome editing”, “neuralink” and the different variants of “autonomous car” – where replaced by “emerging technology”; the basic idea underlying the second approach was to have a measure of the relationship among risk and emerging technologies when considering all the technologies of the case study. Both cases show a peak in 2023. One can then look at the words closest to “emerging technology” to interpret the results; the top 20 per year are reported in Table 2 when not “collapsing” the diverse queries. In the case of “emerging technologies collapsed”, we observed words such as “AI”, “generative”, “vehicle”, “robot”, “automation”, “autopilot”; those words, along with “cybersecurity”, might suggest a possible interpretation of the reasons for a peak of the closeness to risk. A more fine-grained analysis based on the actual documents from that year must then be adopted to confirm the result, and that requires advanced search functionalities to retrieve documents relevant to the task. This is an example of workflow mentioned in Section 2 and that we aim to support.

Another example, relying on TM algorithms, might help to introduce the discussion on controversial issues rooted or related to research on IAR, ML, or AI in general. These indications by PCST scholars might lead to novel research problems to address and result in novel algorithms and paradigms. For example, we considered two newspapers: The New York Times and The Guardian. We used the articles available in TIPS from 1999 to 2022. That resulted in 4,556,415 documents. Then we extracted all the articles answering the query: `(”search engine” OR ”information retrieval” OR ”machine learning” OR ”artificial intelligence”)` in the time interval 1999-2022; 1999 was selected as the starting year because the number of articles in The Guardian seems to be small before that date and we wanted to

Table 2
Top 5 words closest to "Emerging Technology"

2016	2017	2018	2019	2020	2021	2022	2023
innovation	robotics	innovation	AI	AI	innovation	innovation	innovation
advance	AI	AI	blockchain	automation	AI	digitization	diversification
discipline	innovation	automation	robotics	disruptive	automation	domain	domain
advancement	Mana	expertise	tool	innovation	niche	robotics	enabler
opportunity	groundbreaking	tool	automation	tool	expertise	entrepreneurship	infrastructure
entrepreneurial	Hide	disruptive	domain	blockchain	digitisation	tool	entrepreneurship
evolution	consortia	frontier	computing	robotics	robotics	advancement	industrial
meaningful	intelligence	blockchain	intelligence	nanotechnology	entrepreneurship	solution	cybersecurity
insightful	levitation	dynamic	disruptive	intelligence	skill	sustainability	learnings
robotics	artificial	creative	innovation	transformative	domain	cybersecurity	multilateralism
organizational	nanotechnology	solution	algorithm	computing	ecosystem	skilling	solution
frontier	DeepMind	robotics	ML	storytelling	enhancement	cyberspace	organisational
avenue	bioengineering	talent	artificial	digital	disruptive	skill	reform
cosmos	computing	skill	innovate	innovative	enabler	computing	advancement
disruptive	startup	hardware	futuristic	futuristic	skilling	ICT	mobilisation
potential	cloud	idea	Machine	quantum	capability	AI	skill
era	usher	workload	cyberspace	artificial	computing	innovative	upskilling
emergence	field	complexity	cloud	skill	knowledge	creation	complementary
strategy	talent	revolution	societal	Automation	blockchain	learning	automation
immense	nationality	industry	advancement	IoT	advancement	blockchain	biofuel

align the two corpora. We then extracted 30 topics using LDA[10] with 500 iterations and the stop-word list provided by the library. Table 3.2 reports a subset of the topics. For instance, the top documents from topic 8 suggest several concerns related to social media platforms and content, such as misinformation. Topic 14 is focused on preoccupations associated with climate change and the environment. Still, top documents also include how these issues can benefit from AI research results and advancement in the field. Topic 25 concerns recent advances in large language models, and the discussion includes the impact that might have on society. Topic 29 includes a discussion on what AI can bring to Art, but also concerns on the problem of copyright infringement due to some IR and AI technologies; these concerns, for instance, resulted in publishers asking governments to protect their work which is "ingested" by AI-based technologies. Those mentioned in the last paragraphs are only a few examples of the relationship between science, technology, and society.

4. Final remarks

This paper discussed how PCST can benefit from working on Digital Data. We explicitly discussed some issues that need to be addressed, relying on our experience on an ongoing interdisciplinary research project called TIPS and previous works on Historical Newspapers Archives. The solutions to some of these issues are already integrated into the TIPS platform. Section 3 discussed how our effort is helpful for research and teaching activities. Besides the specific activity considered in the paper, we should also point out that the platform has been actively used for several years by Bachelor's, Master's, and Ph.D. students for their theses, which generally focus on specific issues concerning technologies and their impact on society.

A large body of work still needs to be done to provide better support, e.g., implementing other workflows like that described in Section 3.2 or proposed in [24] directly into the platform.

Moreover, some platforms discussed in the paper are useful for research activities; are they

[10] https://mimno.infosci.cornell.edu/jsLDA/

Table 3
Topics extracted from NYT and The Guardian on IAR, AI, and ML.

ID	Top words
1	apple mobile android gt&gt app phone iphone apps google phones
2	students university science education school professor research universities computer online
3	human intelligence computer artificial machine machines humans computers language learning
4	game games video virtual players play player real reality film
5	internet software your users use information web data system computer
6	company said companies percent market business billion investors chief money
7	your home like technology devices voice into smart amazon assistant
8	facebook social users media twitter news content tech youtube company
9	data how used such says could research information about use
11	said online advertising internet ads service companies business web site
12	health patients medical cancer care nhs said doctors patient disease
13	robots robot space human robotics weapons artificial intelligence military robotic
14	energy climate water species said food carbon could change global
15	said data privacy about information public government law had use
16	brain science human his life scientists scientific mind book consciousness
17	google google's microsoft said companies company amazon search european tech
18	said technology software like research computer company a.i companies data
20	jobs workers work job economy automation economic report skills employees
21	cars car vehicles self-driving autonomous uber travel tesla vehicle driving
22	search web site sites information engine online pages your find
23	said covid vaccine health coronavirus were pandemic virus cases had
24	google search engine yahoo google's microsoft users company results internet
25	chatgpt said technology about artificial intelligence google musk use privacy
26	says digital technology business media social guardian such director marketing
27	china chinese government said united states american companies technology china's
28	even about future technology power way might just too much
29	music books art book library digital copyright work artists into

effective also for educational activities? How can we improve them to support teaching better and allow practice on some methodologies/research issues? Should we provide the users with novel search or analysis primitives for better support?

The active participation by experts is a unique opportunity, especially for the evaluation of the effectiveness of IAR, NLP and ML. For this reason, we plan to extend the platform to gather more feedback from the users, e.g., allowing them to specify additional annotations during the labeling procedure, such as notes on why the document was perceived as relevant.

Acknowledgments

The authors would like to thank all the students of the Digital Sociology course in the Master of Communication Strategies at the University of Padova. Without their commitment and participation, the teaching activity reported in this paper would not have been possible.

References

[1] N. Marres, Digital sociology : the reinvention of social research / Noortje Marres, Polity Press, Cambridge Malden, 2017.

[2] G. A. Veltri, Digital social research / Giuseppe A. Veltri, Polity, Cambridge (UK) Medford (MA, USA), 2020.

[3] P. DiMaggio, M. Nag, D. Blei, Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding, Poetics 41 (2013) 570–606. arXiv:9605103.

[4] D. Odijk, B. Burscher, R. Vliegenthart, M. de Rijke, Automatic thematic content analysis: Finding frames in news, in: A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, B. T. Dai (Eds.), Social Informatics, Springer International Publishing, Cham, 2013, pp. 333–345.

[5] M. Bucchi, B. Trench, Science communication research: themes and challenges, Routledge, New York, 2014, pp. 1–14.

[6] P. Catapano, P. Fayard, B. V. Lewenstein, The Public Communication of Science and Technology and International Networking, Springer Netherlands, Dordrecht, 2003, pp. 31–42. doi:10.1007/978-94-017-0801-2_3.

[7] P. Broks, T. Gascoigne, J. Leach, B. V. Lewenstein, L. Massarani, M. Riedlinger, B. Schiele, Communicating science: a global perspective, ANU Press, 2020.

[8] M. Bauer, P. Pansegrau, R. Shukla, The Cultural Authority of Science: Comparing across Europe, Asia, Africa and the Americas, Routledge Studies in Science, Technology and Society, Taylor & Francis, 2019.

[9] P. Magaudda, F. Neresini, Gli studi sociali sulla scienza e la tecnologia, Manuali. Sociologia, Il Mulino, 2020.

[10] E. Pfanzelter, S. Oberbichler, J. Marjanen, P.-C. Langlais, S. Hechl, Digital interfaces of historical newspapers: opportunities, restrictions and recommendations, Journal of Data Mining & Digital Humanities HistoInformatics (2021). URL: https://jdmdh.episciences.org/6121. doi:10.46298/jdmdh.6121.

[11] S. Oberbichler, E. Boroş, A. Doucet, J. Marjanen, E. Pfanzelter, J. Rautiainen, H. Toivonen, M. Tolonen, Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians, Journal of the Association for Information Science and Technology 73 (2022) 225–239. doi:10.1002/asi.24565.

[12] A. Doucet, M. Gasteiner, M. Granroth-Wilding, M. Kaiser, M. Kaukonen, R. Labahn, J.-P. Moreux, G. Muehlberger, E. Pfanzelter, M.-E. Therenty, et al., Newseye: A digital investigator for historical newspapers, in: 15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, 2020.

[13] Impresso Project, 2023. URL: https://impresso-project.ch/.

[14] F. Neresini, Old media and new opportunities for a computational social science on PCST, Journal of Communication 16 (2017).

[15] P. Breucker, J.-P. Cointet, A. Hannud Abdo, G. Orsal, C. de Quatrebarbes, T.-K. Duong, C. Martinez, J. P. Ospina Delgado, L. D. Medina Zuluaga, D. F. Gómez Peña, T. A. Sánchez Castaño, J. Marques da Costa, H. Laglil, L. Villard, M. Barbier, Cortext man-

ager, 2016. URL: https://docs.cortext.net.

[16] I. Flaounas, O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie, N. Cristianini, Noam: News outlets analysis and monitoring system, in: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 1275–1278. doi:10.1145/1989323.1989474.

[17] R. Steinberger, B. Pouliquen, E. van der Goot, An introduction to the Europe Media Monitor family of applications, in: Proceedings of the SIGIR 2009 Workshop on Information Access in a Multilingual World, volume 43, 2009. arXiv:1309.5290.

[18] L. Yang, Y. Wang, D. Dunne, M. Sobolev, M. Naaman, D. Estrin, More than just words, ACM, 2019, pp. 276–284. URL: https://dl.acm.org/doi/10.1145/3289600.3290993. doi:10.1145/3289600.3290993.

[19] J. Qiang, Z. Qian, Y. Li, Y. Yuan, X. Wu, Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 1427–1445. doi:10.1109/TKDE.2020.2992485. arXiv:1904.07695.

[20] J. Qiang, P. Chen, W. Ding, T. Wang, F. Xie, X. Wu, Heterogeneous-Length Text Topic Modeling for Reader-Aware Multi-Document Summarization, ACM Transactions on Knowledge Discovery from Data 13 (2019) 1–21. doi:10.1145/3333030.

[21] A. Cammozzo, E. Di Buccio, F. Neresini, Monitoring technoscientific issues in the news, in: ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, volume 1323 of *Communications in Computer and Information Science*, Springer, 2020, pp. 536–553. doi:10.1007/978-3-030-65965-3_37.

[22] E. Di Buccio, A. Cammozzo, F. Neresini, A. Zanatta, TIPS: search and analytics for social science research, in: L. Tamine, E. Amigó, J. Mothe (Eds.), Proceedings of the 2nd Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2022), Samatan, Gers, France, July 4-7, 2022, volume 3178 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_33.pdf.

[23] E. Di Buccio, A. Lorenzet, M. Melucci, F. Neresini, Unveiling latent states behind social indicators, in: R. Gavaldà, I. Zliobaite, J. Gama (Eds.), Proceedings of the First Workshop on Data Science for Social Good co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Dicovery in Databases, SoGood@ECML-PKDD 2016, Riva del Garda, Italy, September 19, 2016, volume 1831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1831/paper_6.pdf.

[24] F. Neresini, P. Giardullo, E. D. Buccio, A. Cammozzo, Exploring socio-technical future scenarios in the media: the energy transition case in italian daily newspapers, Quality and Quantity 54 (2020) 147–168. doi:10.1007/s11135-019-00947-w.

[25] E. Raff, Jsat: Java statistical analysis tool, a library for machine learning, Journal of Machine Learning Research 18 (2017) 1–5. URL: http://jmlr.org/papers/v18/16-131.html.

[26] H.-F. Yu, F.-L. Huang, C.-J. Lin, Dual coordinate descent methods for logistic regression and maximum entropy models, Machine Learning 85 (2011) 41–75. URL: https://doi.org/10.1007/s10994-010-5221-8. doi:10.1007/s10994-010-5221-8.

[27] D. H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/

S0893-6080(05)80023-1.

[28] E. Di Buccio, F. Neresini, Data from: Research and Teaching Public Communication of Science and Technology on Digital Data, 2024. doi:10.5281/zenodo.10616684.

[29] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[30] V. Di Carlo, F. Bianchi, M. Palmonari, Training temporal word embeddings with a compass, in: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 2019, pp. 6326–6334. doi:10.1609/aaai.v33i01.33016326. arXiv:1906.02376.