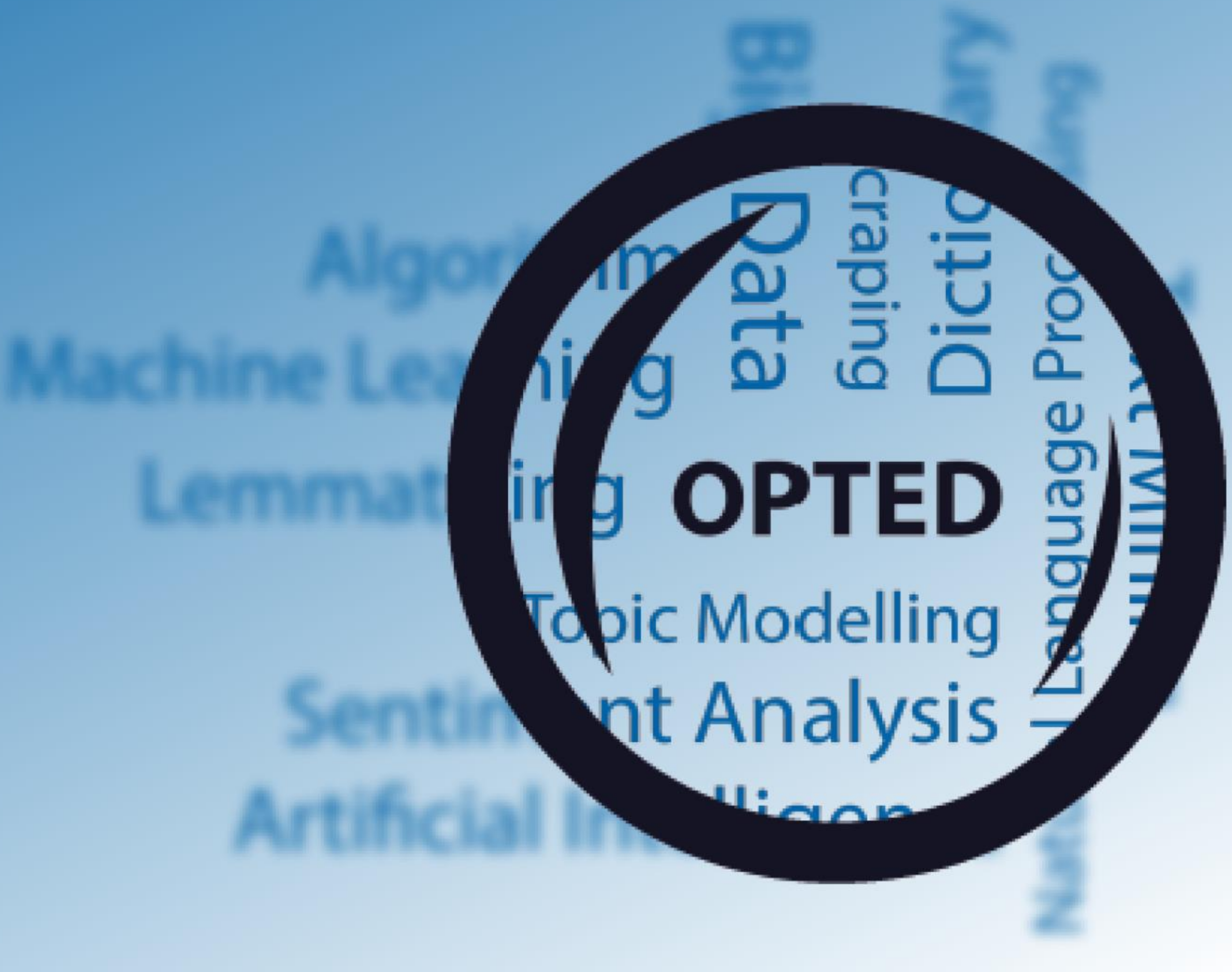
OPTED

A review of citizen-produced political text (CPPT) across time and languages: Data, tools, methodologies and theories

Shota Gelovani, Bente Kalsnes, Karolina Koc-Michalska, Yannis Theocharis



Dislaimer

This project has received funding from the European Union’s Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. The document reflects only the authors’ views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

A review of citizen-produced political text (CPPT) across time and languages: Data, tools, methodologies and theories

Deliverable D2.2

Authors: Shota Gelovani[1], Bente Kalsnes[2], Karolina Koc-Michalska[3], Yannis Theocharis[1]

[1] Technical University of Munich, Germany
[2] Kristiania University College, Norway
[3] Audencia Business School, France

Due date: September 2021

1 A review of citizen-produced political text (CPPT) across time and languages: Data, tools, methodologies and theories

This review was initially envisaged as a systematic review focused on theoretical approaches used in prior literature to contextualize CPPT. Upon the exhaustive review of the literature across time and languages carried out as part of deliverable 2.1, it became clear that such a task would be less interesting and valuable for this project and beyond the scope of WP2 and OPTED goals. This was for several reasons.

The first reason was practical and concerned with the way that the data had to be collected. The creation of a codebook of such a magnitude and multilingual nature, and which aimed at uncovering how CPPT is utilized in existing scholarship, relies principally on quantitative indicators that have to be manually extracted and classified by human coders. This is a very time-consuming and challenging process that requires several rounds of training human coders and attention to detail. Ultimately, as the coding categories refer to relatively straightforward constructs that could be eventually agreed upon (e.g. which method, language or software did the study use (please see Coding book in D2.1)), this coding exercise did not face more significant challenges than those faced by other researchers engaged in this type of manual classification. The number of relevant manuscripts about CPPT, identified through a combination of automatized webscraping on Google Scholar, manual search via local engines, and human coding is 3,260. Coding *theoretical* approaches would therefore be challenging mostly because it would require more in-depth qualitative coding approach, which was impossible within the scope of the manuscripts classified as relevant.

This relates to the second reason why the scope of the second deliverable had to be adjusted. The kind of content analysis we carried out proved to be unsuitable as a method for coding theoretical approaches of a multifaceted concept such as CPPT (a conceptual innovation developed by this consortium and thus a construct not yet well-established in the existing literature). Not only it is extremely difficult to extract theoretical conceptualization from existing literature using CPPT because such conceptualizations are rarely mentioned explicitly by authors (the lack of theory in this line of scholarship has been pointed out by scholars, i.e. Salganik, 2007). But even when the authors do theoretically contextualize their work, it is difficult for coders to dig out mentions to "theories" and then agree on the specific theoretical or conceptual constructs being used. Concretely with our data, in some (few) cases authors explicitly mentioned using, e.g. the "spiral of silence" theory in studies where CPPT was used as a data source, others simply referred to theoretically plausible relations between various concepts, rendering it unlikely to accurately code *the* conceptual and theoretical framework.

Our solution to this challenge was twofold. Firstly, we use this deliverable to present a host of evidence related to how the use of CPPT by scholars has evolved over the last decades and across languages in manifold way – *including theory*, to the extent we could uncover the theoretical foundations of studies. Secondly, using unsupervised text analysis methods, we extracted content related to theoretical approaches by selecting specific "theory" – related keywords, including the word theory itself. The results of this exercise, which are reported here, confirm the futility of the initially envisaged review of theoretical approaches, as mentions to "theory" or "theoretical" approaches are rather scarce in our corpus of thousands of articles.

In the chapters that follow we discuss the use of CPPT across time and languages, main sources of CPPT (social media, forums, blogs etc.), as well as tools and methods for collecting and analyzing them. We conclude by contextualization of CPPT.

2 CPPT use across time and languages

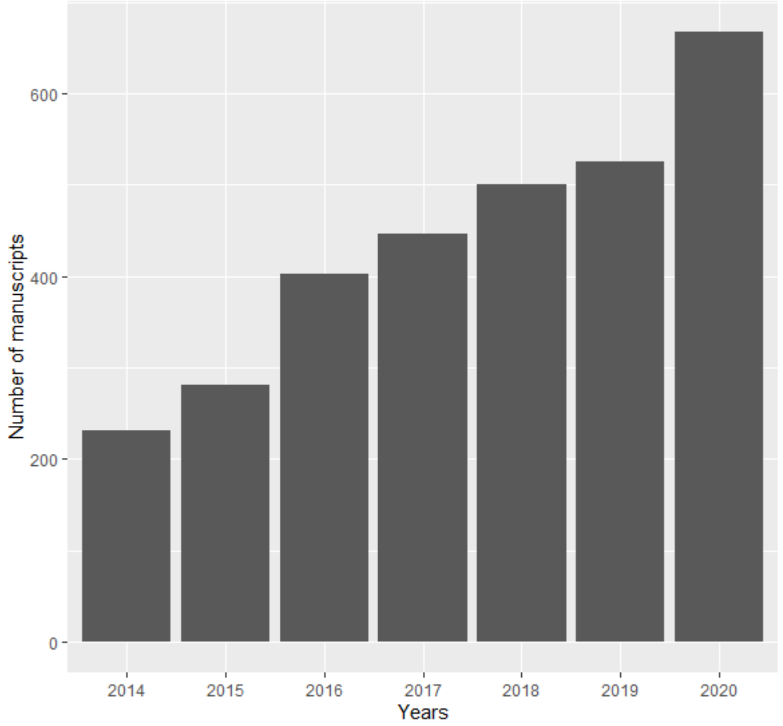
We developed 37 search queries to sufficiently capture academic manuscripts about CPPT. English manuscripts (only academic publications in peer-reviewed journals, please see detailed description of the methodological approach in deliverable 2.1) were scraped via Publish or Perish software from Google Scholar. The non-English manuscripts were manually coded by research assistants into the dataset by entering translated (and adapted to local context) search queries in various local search engines.

Automatically scraped English manuscripts went through a superficial, metadata-based filtering (I stage) process that removed the following manuscripts:

- Published in a journal from a completely irrelevant discipline (e.g. health, geography, law studies, etc.)
- Books, book chapters, book reviews, theses, non-academic articles, working papers, citations
- Manuscripts only uploaded to repositories (university repositories, Researchgate, arXiv, academia.edu, and various other repositories)
- Duplicates

The filtering (I stage) left us with a total of 6,040 **likely** relevant articles in the period of 2014-2020 for manuscripts (articles) published in English. Further filtering (stage II) took place during the coding stage. Coders could label the manuscript as irrelevant based on content (e.g. if the manuscript did not contain CPPT, there were no data, were not political etc.) The coding of both English and non-English articles resulted in 3,260 relevant manuscripts published from 2014 to 2020. A time series of CPPT-related manuscripts is given on Figure 1. The number of coded relevant articles in the dataset is on the rise. As the new means of communication technology emerge, so does the scientific interest in them.

Figure 1 NUMBER OF RELEVANT MANUSCRIPTS IN THE CPPT DATASET BY YEAR OF PUBLISHING



A detailed list of languages in which the articles were written is presented in Table 1. Out of 3,260 relevant articles coded between 2014 and 2020, 2,109 were published in English and 1,151 were published in other languages than English.

Table 1 NUMBER OF CODED MANUSCRIPTS BY LANGUAGES IN WHICH THEY ARE WRITTEN

	English	French	German	Italian	Portuguese	Spanish	Polish	Norwegian	Swedish	**Total**
Journal articles	2109	346	65	63	44	66	56	18	0	**2765**
Books	0	1	6	1	0	0	7	0	1	**16**
Book	0	0	49	0	8	5	15	2	0	**79**

chapters										
Conference proceedings	0	20	2	8	10	11	3	0	0	**55**
Reports	0	1	11	0	2	0	1	5	2	**22**
Theses or dissertation	0	50	82	27	37	13	3	35	5	**253**
Working papers	0	7	20	0	0	1	1	0	0	**29**
Total	**2109**	**445**	**235**	**120**	**101**	**96**	**86**	**60**	**8**	**3260**

The CPPT written in those languages are covering different geographical areas. Europe (in general) and North America are covered by 2,153 articles, together accounting for two-thirds of the entire dataset.

Table 2 **REGION OF THE ORIGIN OF THE CPPT DATA**

Region	# of manuscripts (% of total)
Europe	1455 (45%)*
North America	793 (24%)
Asia	466 (14%)
South America	200 (6%)
Middle East and North Africa	195 (6%)
Does not specify	174 (5%)**
Sub Saharan Africa	137 (4%)
Australia and Oceania	103 (3%)
Central America	20 (1%)

**Regions are not mutually exclusive and several regions could have been covered in one manuscript*
***174 manuscripts are analyzing the data without identifying the region, possibly employing big-data sets not concentrating on specific place but rather inter-cultural exchange*

We find similar results if we look at top 10 countries covered by the manuscripts. There is a definite overrepresentation of studies covering the US, but we also find China high on this list. Canada and Brazil are two countries not originally covered by our study, but are appearing on the list with three native languages employed in the study (Canada: English and French, and Brazil: Portuguese). In total, CPPT from 154 countries were analyzed in the manuscripts.

Table 3 **TOP 10 COUNTRIES COVERED BY THE MANUSCRIPTS**

Country	# of manuscripts (% of total)
US	650 (20%)
France	285 (9%)
Germany	239 (7%)
UK	208 (6%)
Italy	160 (5%)
China	142 (4%)
Brazil	133 (4%)
Canada	124 (4%)
Poland	115 (4%)
Spain	106 (3%)

Table 4 indicates the top 10 languages in which the data sets containing CPPT are constructed;

again – this is not an exclusive count as one manuscript could employ data in several languages. Even if the results from Table 1 are biased by our selection of the countries/languages covered in the study, it is visible that English is predominant among the analyzed languages. Within the top 10, English based data sets (1,517) are employed almost equally to all other languages analyzed (1,673). Two of the languages in top 10, Chinese and Arabic, are categorized but not covered by the language-specific search within our study. Those are the two most popular languages that are analyzed by researchers and published in non-language-specific journals. In total, CPPT from 134 languages and dialects were analyzed in the relevant manuscripts.

Table 4 **TOP 10 LANGUAGES COVERED BY THE MANUSCRIPTS**

Language	# of manuscripts (% of total)
English	1517 (47%)
French	443 (14%)
German	313 (10%)
Spanish	197 (6%)
Italian	171 (5%)
Portuguese	139 (4%)
Chinese	126 (4%)
Polish	114 (4%)
Arabic	94 (3%)
Norwegian	76 (2%)

3 CPPT as a data source

Different sources of CPPT were included in the database covering social media (Facebook comments, Facebook posts, original tweets, Twitter comments, Instagram posts etc.), other websites (blogs, forums, political deliberation websites etc.), and offline CPPT (letters to the editor, citizen opinions in the newspapers). There was a total of 27 options in the codebook. Again, the coders were allowed to select several options. It is visible that social media dominate as a source of the CPPT data. Facebook is the main source of CPPT – a total of 1,096 manuscripts use Facebook data (posts and comments) as CPPT, which accounts for 34% of all manuscripts. It is closely followed by Twitter (includes original tweets and comments, 780 manuscripts, 24% of all manuscripts), and online newspapers (580, 18% of all manuscripts). Nevertheless, we have to underline that the data cover the period from 2014 to 2020, so it is possible that if we would cover a more extensive period of time, we would find more results based on offline sources or internet enabled communication (e.g. forums or blogs) or organizational/institutional online sources. The manuscripts based on CPPT data originating from any kind of social media platforms constitute six (counting YouTube as well) out of top ten sources. Additionally, we find CPPT data from online newspapers, blogs, forums and political/deliberation websites.

Table 5 **THE TOP 10 MOST FREQUENT SOURCES OF CPPT DATA**

Sources	# of manuscripts (% of total)
Facebook posts	855 (26%)
Facebook comments	789 (24%)
Original tweets	711 (22%)
Online newspapers	580 (18%)
Retweets or replies	420 (13%)
Blogs	387 (12%)
Forums	366 (11%)
Original YouTube videos	200 (6%)

YouTube comments	143 (4%)
Political/deliberation websites[1]	121 (4%)

The coders were asked to indicate a URL of the dataset if it was available. Out of 3,260 relevant articles, only 43 (1.3% of all manuscripts) indicated a URL to their dataset in the form of supplementary materials to the article (5 manuscripts), GitHub repositories (4 manuscripts), YouTube links (4), blogs (4), other repositories (3), as well as Dropbox (3) or Reddit thread links (2). It needs to be noted, however, that even though the manuscripts do not explicitly indicate the URL to the dataset, many of them mention the online sources (blogs, websites, social networking sites) where they accessed the data. Some of them may have only restricted access or be subject to privacy concerns (social networking sites data), while others, such as blogs or forum threads, are often mentioned in the manuscripts at face value, without a URL. Hence, the number of available datasets should be estimated as more than 43. Nevertheless, this finding adds to the serious concerns about the availability of the data, the reproducibility of the research. This finding demonstrates that Open Science principles[2] are hardly followed in the existing research material about CCPT.

4 CPPT collection and analysis tools

Table 6 indicates the five most important approaches to collect CPPT data in the research literature. The vast majority of the data sets are collected by hand (copy-paste) by the researchers. However, there is a growing number of data acquired from professional companies or downloaded via computer programs written specifically by the authors of the research article.

Table 6 **THE MOST POPULAR METHODS OF COLLECTING THE CPPT DATA (NUMBER OF MANUSCRIPTS)**

CPPT collection methods	Description	# of manuscripts (% of total)
Self-copy-paste	data copied directly from the source without an intermediary software, ‘by hand’	1552 (48%)
Company/bought	data purchased or obtained from a third-party app/software/company, i.e., there is quality in the data, but no control over it	576 (18%)
Dictionaries/keyword searches	data collected by searching dictionaries/repositories/websites/apps for keywords	497 (15%)
Interviews	data collected via interviews	495 (15%)
Self-written program	data collected using a self-written program	234 (7%)

The number of reported software used to collect CPPT data was low. The most popular program used for scraping, collecting, or downloading the CPPT data included Twitter API (used in 124 manuscripts), Netvizz (42), Qualtrics (36), Facebook Graph API (34), NCapture (14), Amazon Mechanical Turk (11), Facepager (7), Tweet Archivist (7), YouTube API (7), and Netlytic (6). All of the top 10 software employed for scraping the data are open source, free of charge software. However, six of them are embedded within the social media platforms. Such a situation confirms the issues with the data credibility and availability (please see for example APNews 2021[3]). Moreover, non-English

[1] Includes petition websites, websites for communicating with politicians, and other websites that facilitate citizens’ political participation.

[2] https://www-nature-com.audenciagroup.idm.oclc.org/articles/s41597-021-00981-0

[3] https://apnews.com/article/technology-business-5d3021ed9f193bf249c3af158b128d18 [Last accessed on September 23, 2021]

articles only accounted for 24% of the manuscripts that used the CPPT collection software mentioned in the top 10.

It seems crucial for the future research to produce more exchange and build an infrastructure helping researchers to cope with the huge amount of data available, but at the same time not really publicly available for continuation or comparability within other project. One of the main goals of OPTED is to set a promises for such infrastructure building.

5 CPPT analysis methods

Overall, 1,661 manuscripts employed qualitative methods of text analysis, 869 employed quantitative or computational methods, whereas 730 used a mixture of the two aforementioned categories. The ten most widely used quantitative/computational and qualitative methods are presented in Table 7.

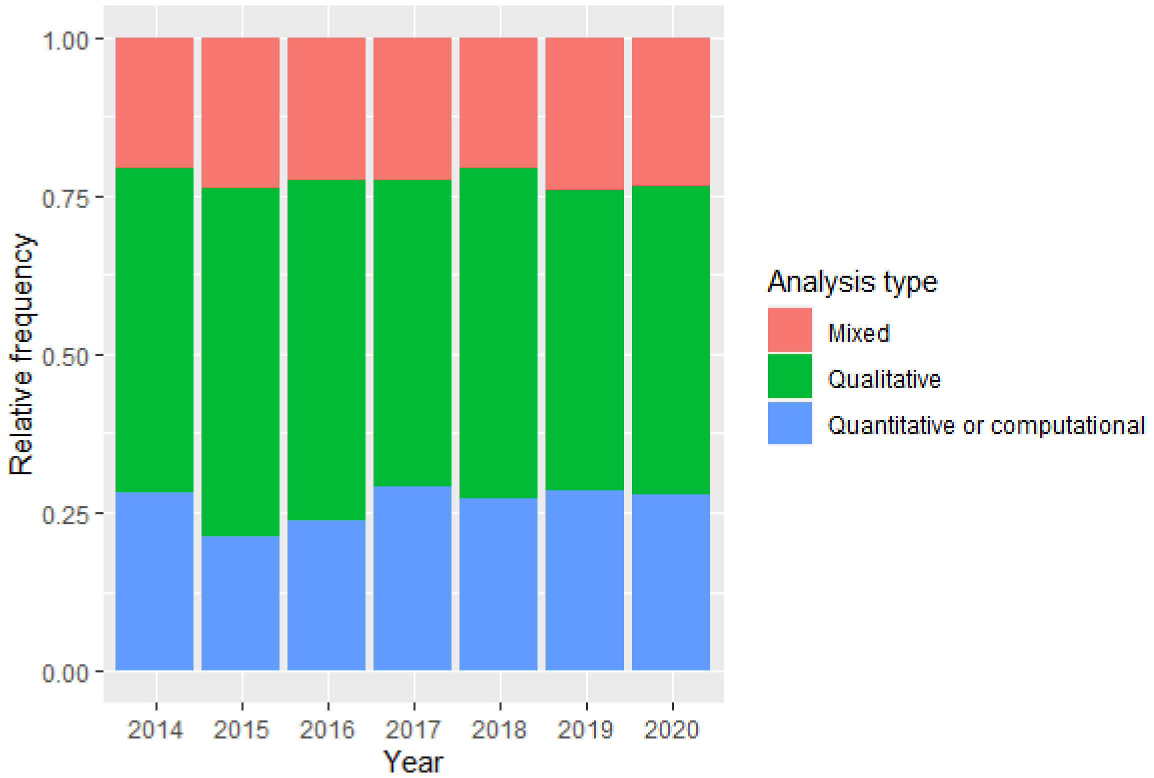
Table 7 **TOP 10 QUANTITATIVE/COMPUTATIONAL AND QUALITATIVE METHODS USED (METHODS AND THE CORRESPONDING NUMBER OF MANUSCRIPTS)**

Quantitative/Computational Text Analysis Methods		Qualitative Text Analysis Methods	
Quantitative content analysis	782	Qualitative content analysis[4]	1105
Text statistics	658	Discourse analysis	847
Manual coding	373	Thematic qualitative text analysis	555
Sentiment scoring	259	Interview	440
Dictionaries keyword searches	152	Observation	395
Topic models or text clustering tools	150	Evaluative qualitative text analysis	220
Automated extraction	143	Type building text analysis	140
Semantic network tools	142	Grounded theory	89
Text similarity scoring	97	Focus group	55
Supervised machine learning	91	Survey	13

Throughout the analyzed period, the use of method types has been stable. The average share of quantitative or computational text analysis methods was 26.5 percent, mixed methods – 26.6 percent, whereas the qualitative text analysis methods were used for approximately a half (51%) of all articles between 2014 and 2020 (Figure 2).

Figure 2 **ANALYSIS TYPES BY YEAR, 2011-2020.**

[4] Qualitative content analysis was understood by the RAs as the default category if the qualitative text analysis method was not specified in the study. However, the methods in the codebook were not mutually exclusive, so the coders could pick more than one.

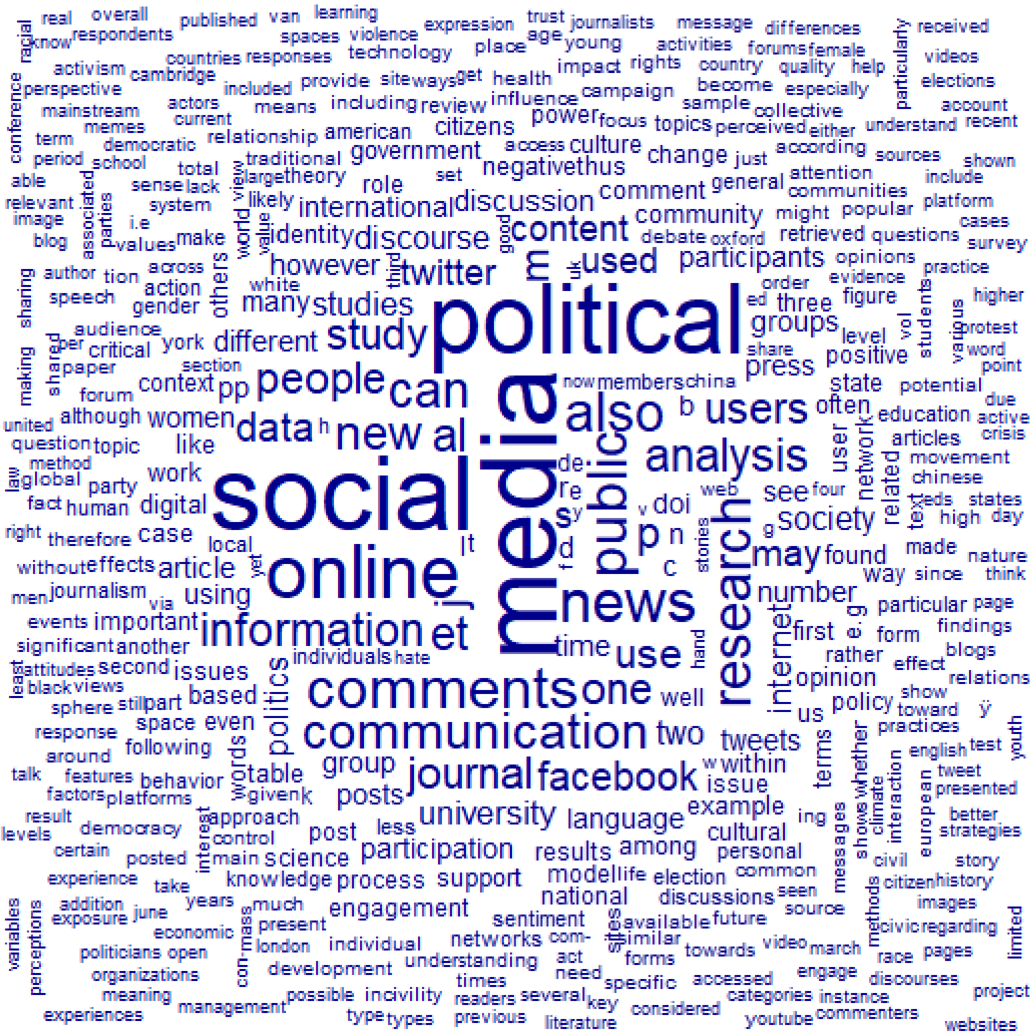


As with the CPPT collection software, there were only a few reports of software used to analyze CPPT. These included both programming languages and standalone software. The most popular CPPT analysis software included Nvivo (102), R and its packages (57), Python and its libraries/demos (45), Microsoft Excel (36), Gephi (29), Iramuteq (25), Atlas.ti (18), NodeXL (16), MAXQDA (15), and Sentistrength (9). One of the important goals that OPTED project aims to address is the issue of low transparency of data analysis which yields for creating a data infrastructure, where the standards and usability of coding software, dictionaries and other analytical platforms would be laid out.

6 Contextualizing CPPT in current research

The task of WP2 was to map the location of CPPT in the scientific literature and research. One way to address this problem is to identify the disciplines where the articles about CPPT belong to. Broadly, they belong to communication or information research, mostly to the strand of communication studies that focus on the new means of communication such as the Internet and social media. The studies of CPPT mainly look at social media data, where citizens express their political opinions or enter discussions about politics via posting, commenting, or messaging (Figure 3). This may have been conditioned by the choice of search queries as well – as mentioned above, 26 out of 37 queries included a name of a social networking website (Facebook, Twitter, Reddit, TikTok, Telegram, WhatsApp, YouTube), while identifying studies that dealt with offline CPPT was very difficult. Other possibility is that there was a shift in research focus after the development of social media and its popularization within political strategies. Such a change from offline or online-enabled CPPT research could have been exercised and published before the time scope of this project (starting in 2014).

Figure 3 **WORD CLOUD OF THE TERMS USED IN THE CODED ARTICLES.**



CPPT as data are located in numerous international journals within different fields. However, since our research was limited to the manuscripts employing the data, theoretical manuscripts, and as a consequence more theory-oriented journals are not present in the top 10 list of scientific journals. As Table 8 indicates, the CPPT based manuscripts have the potential to be published in the top-tier international journals, with most manuscripts being published in New Media and Society (ranked 5th in Communication journals[5]), Information, Communication & Society (ranked 10th) or Social Media+Society (ranked 17th). Besides the English language journals, in top 10 we find also two French language journals. In total, there were 2,765 articles (85% of all manuscripts) published in 1,668 journals.

Table 8	MOST RELEVANT SCIENTIFIC JOURNALS COVERING RESEARCH ON CPPT
Source	# of manuscripts (% of all journals)
New Media & Society	81 (5%)
International Journal of Communication	73 (4%)
Information, Communication & Society	60 (4%)
Social Media+ Society	52 (3%)

[5] https://www.scimagojr.com/journalrank.php?area=3300&category=3315 [accessed September 2021]

Computers in Human Behavior	47 (3%)
Réseaux	29 (2%)
Social Science Computer Review	26 (2%)
Journalism	23 (1%)
Argumentation et Analyse du Discours	21 (1%)
Journal of Information Technology & Politics	21 (1%)

The inductive approach towards theory contextualization produced interesting results. As discussed in section 1, the theoretical approaches in the coded CPPT articles were more challenging to quantify in the face of a large corpus of articles. It was also challenging to create a common coding framework for theories used to study CPPT in the face of the absence of any prior theoretical contextualization on CPPT. Therefore, theories used in the manuscripts were identified *after* the coding was done and the final dataset was available. Table 9 presents the overall co-occurrences with the word 'theory' within the entire corpus of the English based manuscripts, while Table 10 presents it for articles in all other languages.

At first glance, it appears that prominent and widely used theories in the field of media and communication, such as social identity theory or spiral of silence theory, but also qualitative methodological approach like grounded theory, were often mentioned in the texts. One should however be cautious here. The inductive keyword-based method used here does not guarantee that these are mentions to theories utilized in the scholarly articles. It might well be, for example, that the authors referred to particular theories in some cases, as illustrated by the majority of theories in Table 9, but the word theory could also be accompanied by broader categories, such as communication or democratic. Conspiracy theory, which is qualitatively different from other theories in that it does not stand for a specific theory, features in six out of nine languages and is the most widespread co-occurrence of all. Theory of communicative action features in five languages, while communication theory is mentioned in four.

Table 9 **CO-OCCURRENCES FOR THE WORD 'THEORY' IN ENGLISH ARTICLES**

Co-occurring term with 'theory'	# of occurrences
Communication	381
Conspiracy	172
Discourse	141
Social Identity	109
Feminist	105
Spiral of Silence	94
Democratic[6]	87
Framing	69
Communicative action	53
Critical race	53

Table 10 **CO-OCCURRENCES FOR THE WORD 'THEORY' IN LOCAL ARTICLES. NUMBERS IN PARENTHESES INDICATE THE NUMBER OF OCCURRENCES.**

French	German	Italian	Norwegian	Polish	Portuguese	Spanish	Swedish
(Théorie du) genre (182)	Theorie des kommunikativen handelns (54)	(Teoria del) Complotto (22)	Retoriske situasjon[7] (18)	(Teoria) Informacji (31)	(Teoria da) Comunicação (20)	(Teoría del) actor red (37)	Social kontroll (14)

[6] Includes 26 occurences of *Theory of deliberative democracy*.

[7] Includes 4 occurences of *Retorisk teori*.

Complot (102)	Demokratietheorie (95)	Two step flow of communication (8)	Framingteori (17)	Dyskursu (19)	Conspiração (19)	Comunicación humana (21)	Habermas teori om borgerlig offentlighet (12)
Communication (38)	Verschwörungstheorie (60)	Relazioni sociali (6)	Diskursteori (14)	Masowej komunikacji (11)	Memética (16)	Democracia[8] (9)	Subaktivism (7)
Argumentation (31)	Diskurstheorie (38)	Influenza selettiva (5)	Subkulturteori (10)	Konfliktu (11)	Representações sociais (14)	Feminista (6)	Deliberativ demokrati (6)
Agir communicationnel (23)	Systemtheorie (37)	Agire comunicativo (4)	Feiministisk (10)	Spiskowych (10)	Discurso (12)	Inteligencia afectiva (6)	Politisering av religion (4)
Réseau (21)	Sprechakttheorie (31)	Agenda setting (4)	Imagegjenopprettelse (9)	Agenda-setting (9)	Relações internacionais (7)	Argumentación (6)	Dagordningsteori (3)
Fonctions de croyance (17)	Medientheorie[9] (28)	(Modello teorico-operativo di) incivility (4)	Offentlighetsteori (8)	Gier (8)	Processo politico (5)	Agenda setting (5)	Medialisering (2)
Action collective (16)	Praxistheorie (23)	Eventi mediali (3)	Demokratiteori (8)	Aktora-sieci (7)	Humor Criptografado (5)	Comunicación (5)	Marknadsföringsteori (1)
Utilisations et satisfactions (16)	Sprachtheorie (17)	(Teoria weberiana della) burocrazia (3)	Feltteori (7)	Demokracji (6)	(Teorias dos) movimentos sociais (4)	Encuadre (5)	Maktteori (1)
Discours (16)	Spieltheorie (15)	Identità sociale (3)	Queerteori (6)	Systemów (5)	Ação comunicativa (3)	Urdimbre comunicativa (4)	Diskursteori (1)

7 Conclusion

The presented deliverable produces some interesting outcomes. One of the most striking finding is the scarcity of the publicly available datasets: Only 43 out of 3,260 manuscripts indicate a link to an available dataset. Most of the tools used to collect or analyze CPPT are free or open-source, however often provided by the large IT industries often also owning the discussion platforms.

[8] Includes 4 occurences of *Teoría de la democracia deliberative.*

[9] Includes 7 occurrences of *Theorie der Medien.*

As could have been expected, most of the manuscripts are published in English, however they cover research in other languages (data based on CPPT produced in local languages besides English). Next most popular languages (within our sample of non-English publications) are French and German. The innovative methods (for example, mixed-approach, AI-based content analysis, unsupervised/automatic methods) are still quite rare and vast majority of the work is based on small-n qualitative methods.

We believe that those and other more detailed findings from the deliverable 2.2 show clearly the need of building a common infrastructure, help construct the cooperation among researchers and run better, well-coordinated studies in the future that would allow for more strict data control and would provide a better understanding of the analytical tools employed. Those are the main goals of OPTED project.

Lastly, the next deliverables from WP2 are to concentrate on challenges the research community is facing while working with CPPT data and solutions that hopefully can be found to overcome those problems.