D1.3: Visualisations of key emerging technologies and social issues (Part II)

Work Package	**WP1: Topic identification**
Task	**Task 1.2 Continuous identification of new technologies and trends**
Due Date	**31/8/2020**
Submission Date	**31/8/2020**
Deliverable Lead	**DELab**
Deliverable Support	
Dissemination Level	**Public**
Nature	**Report**
Authors	**DELab: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, Maciej Wilamowski**
Reviewers	**Louis Stupple-Harris**
Status	**Final**

Acknowledgement: This report is funded under the EC H2020 project NGI Forward, Grant Agreement 825652.

Table of Contents

NGI FORWARD

EXPLORATIONS IN NEXT GENERATION INTERNET

View interactive presentations to see the visualisations and overview of our main findings:

- **the first part of the report covering years 2016-2020 can be found at: https://fwd.delabapps.eu/,**
- **the second part dedicated to the COVID-19 pandemic analysis is available at: https://covid19.delabapps.eu/.**

1. Introduction

Our daily lives are disrupted by the ongoing COVID-19 pandemic. The crisis has changed the way we work, study, travel and spend our free time. The struggle of governments, institutions and companies to keep track of the situation has brought to the fore crucial discussions on the collection and use of data. In the meantime, existing social challenges, such as climate change or the spread of misinformation, persist.

Recognising the relationship between technology, social challenges and current events is, therefore, more important than ever. In the spirit of the ongoing challenges, our team has been monitoring the tech world to gather and organise insights on the most relevant trends. Equipped with various text-mining tools, we have been trying to make sense of the information maze surrounding us.

This report aims to summarise our conclusions from analysing the most important developments in two time periods: the years 2016 to 2020, and the six months between January and June 2020. By dividing our analysis into two parts, we can focus more on the long-term challenges, and also better understand the current crisis.

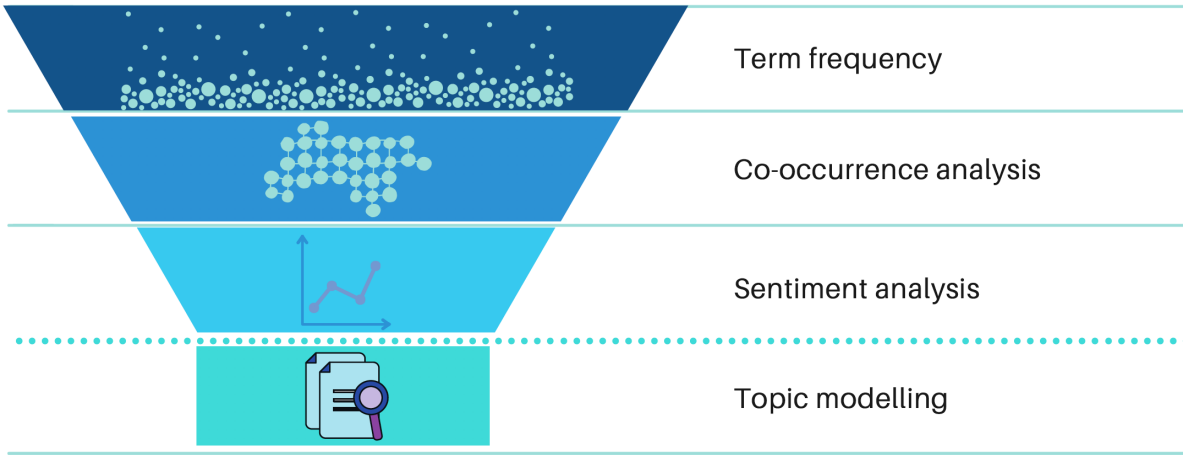
In the first part of the study, we continue the multi-stage topic identification process used in our previous research deliverables. Building on the insights from 2019 (summarised in reports D1.2 and D1.9), we examine trends in news articles and prepare deep dives for 8 key NGI topics:

- Trustworthy Information
- Blockchain and cryptocurrencies
- Online Privacy
- Sustainability and Climate Crisis
- Safer Online Environments
- Democracy
- Market Competition
- Ethical AI

In the second section, we focus on the COVID-19 crisis. We analyse trends in tech news, open-source projects on GitHub and Reddit discussions. Since April 2020, we have regularly released new insights in an online presentation. In the report, we summarise key conclusions based on up to date data.

The method in the study is based on our prior works (D1.2, D1.5). For the identification of trends and further exploration of key issues, we use the combination of various methods. First, we identify emerging topics based on the analysis of term frequencies. The terms with the greatest increase of frequency over time can be filtered using regression analysis. The identified terms serve as input for further analysis. The connections between emerging social issues (such as fake news or privacy) and technologies are explored using co-occurrence and sentiment analysis techniques. The co-occurrence values are calculated between selected emerging trending topics, pinpointing pairs of terms that are most frequently mentioned together (such as “hate speech” and “machine learning”). Sentiment analysis is then performed to track the public perception of issues and identify the positive and negative news stories related to a selected topic. To further verify our results, topic modelling is conducted using Latent Dirichlet Allocation (LDA), a complementary approach to our topic identification strategy. LDA shows which terms define key topics across the text corpus, providing information on wide topics and co-occurrences. The pipeline of the analysis is presented in Figure 1.

Figure 1 Pipeline of the analysis



In the second part of the study, we also experiment with other techniques to map and visualise the topics in news articles. We iterate on the methods presented in D1.5 (LDA and t-SNE), and also explore dynamic topic modelling.

Following the Introduction, we briefly introduce the sources and methods. Readers interested in the results and analyses can directly skip to Section 4. Topic identification, Section 5. Deep dives and Section 6. Covid-19 analysis.

The codes to replicate our analyses are published online at Gitlab (https://gitlab.com/enginehouse). The presented results are available in the form of interactive visualisations at https://fwd.delabapps.eu/ and https://covid19.delabapps.eu/. The raw results are also publicly available for further analysis (https://zenodo.org/communities/ngi_forward/).

2. Sources

The forecasting exercise described here is based on the analysis of information stemming from four important stakeholder communities: journalists covering technology news, academia, social media (Reddit) and open source (GitHub). The latter two are used only in the COVID section of this research.
Online news websites reporting on the tech world provide us rich data on the topics driving public discussions, while the dynamics in academic research reveal areas gaining a lot of traction, indicating that they are likely to be implemented in the future. The main methodology combines these two perspectives, providing a more holistic insight into the development of Internet technologies. Additionally, during the coronavirus pandemic, a rapidly changing discussion landscape meant that particular topics needed to be examined using social media sources. Cooperation and social attitudes are better gauged by information stemming from the crowd. These sources are represented by a news aggregator and discussion forum Reddit, and by GitHub – a software development hosting provider and an open-source collaboration hub.

The news data has been collected for a period of 48 months (between 2016-01-01 and 2019-12-31) and separately for the pandemic period between January and June 2020 (2020-01-01 to 2020-06-30).

Figure 2 presents the 14 selected online media sources, including the number of collected articles and country of origin. Altogether, we collected 250,000 articles for the first period and over 30,000 for the COVID-19 analysis.

In the case of working papers, works from two repositories are collected: arXiv (STEM sciences) and SSRN (social sciences).
ArXiv is owned and operated by Cornell University. Originally created as a physics archive, the arXiv repository's remit has expanded and currently covers a wide range of sciences, including computer science. The computer category within the arXiv repository deals with topics such as Artificial Intelligence, Computation and Language, Cryptography and Security, Data Structures and Algorithms, Human-Computer Interaction, Information Retrieval, Networking and Internet Architecture.
SSRN (The Social Science Research Network) is considered to be one of the leading social science and humanities online repositories. SSRN is owned by Elsevier.

The overall number of the analysed working papers on arXiv reach 155,000 (and additional 34,000 during COVID-19), while working papers on social sciences include over 25,000 (and 4,000 for COVID-19).

Reddit, which calls itself "the front page of the internet", is a news aggregator and a discussion forum. Based on Alexa, it is the third largest website in Sweden, Finland, Denmark and the Netherlands, only behind Google and Youtube; in some other countries (like Belgium, Germany and Croatia) it is also in the top ten.

The website is split into subreddits – forums with a common topic, with its own moderators, userbase, and often even a distinguishable culture. The importance of Reddit as a leading website for discussions cannot be understated. The subreddit r/Coronavirus is one of the most active communities, with over 2 million subscribers and posts reaching tens of thousands of upvotes every day. Apart from the main discussion hub, the topic is discussed in a number of smaller communities.

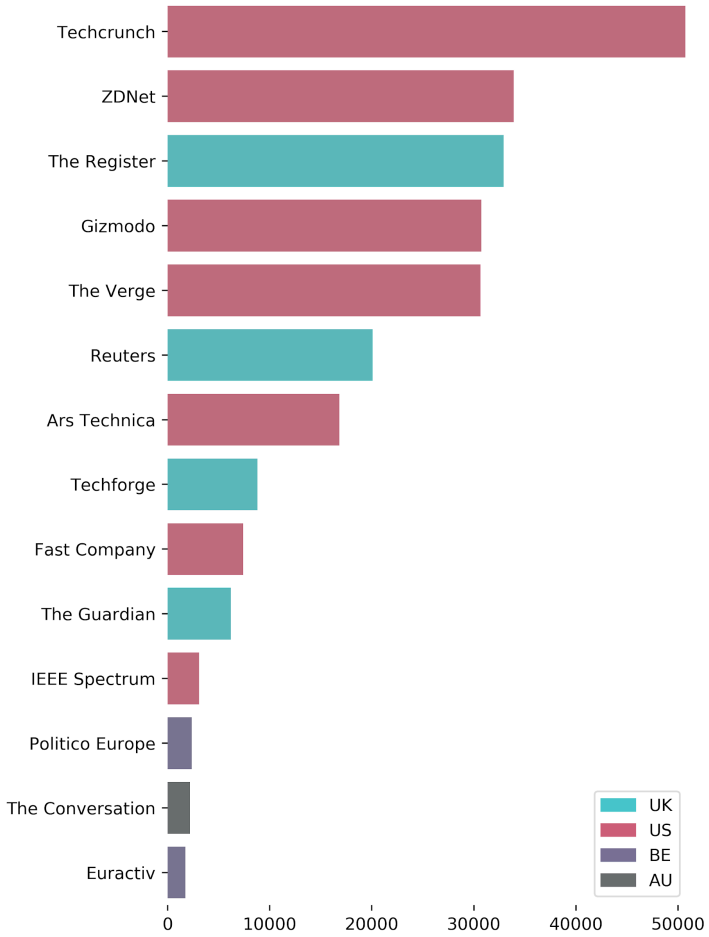
Data has been downloaded using the Pushshift API. It is not the official API, but its filtering capabilities and speed make it preferable to the Reddit API. As the GitHub documentation page says, “The pushshift.io Reddit API was designed and created by the /r/datasets mod team to help provide enhanced functionality and search capabilities for searching Reddit comments and submissions.” A given number of comments or submissions (500 at the beginning of the analysis, the limit has been later reduced to 100) has been downloaded for each topic in a single request; shortly after a request has been successful, another – with an applicable time limit – has been sent in accordance with the limits imposed by the API. Comments and submissions have been downloaded from 2020-01-01, and the date limit for charts is 2020-07-10. In one case (sentiment analysis), due to the low number of comments and seemingly random results before the pandemic regarding “remote work”, the analysis starts from March.

GitHub is the most commonly used hosting platform for open-source developers, using it as a collaborative version control system. Despite its recent acquisition by Microsoft, which has a problematic history with open-source projects,[1] it has not been abandoned – quite the opposite, recent months have seen a stable growth in the number of users, with 10 million new users in 2019.[2]

The data source for our analysis has been sourced from GitHub itself[3]: the platform provides versioned dumps of all repositories connected to the coronavirus. The files used were .json data from between 2020-01-20 and 2020-07-13. Data contains the repository URL and description, topics, information about owner, creation date, primary programming language, license, and counts of actions (forks, contributions and commits, among other things).

[1] https://www.theregister.com/2001/06/02/ballmer_linux_is_a_cancer/
[2] https://octoverse.github.com/
[3] https://github.com/github/covid-19-repo-data

Figure 2 Number of media articles per source and country of origin



3. Methods

Our sequential text-mining method contains 4 steps:

1. Topic identification
2. Co-occurrence analysis
3. Sentiment analysis
4. Topic modelling

The description of these methods originates from our paper in review (Gyódi, Kristóf and Nawaro, Łukasz and Paliński, Michał and Wilamowski, Maciej: Informing Policy with Text Mining: Technological Change and Social Challenges, *Technological Forecasting and Social Change*, in-review).

Following the introduction to these methods, we also provide a brief summary of other methods used in the study (t-SNE, dynamic topic modelling). For more information on dimensionality reduction techniques, we refer to the D1.5 report (available online at https://fwd.delabapps.eu/topic_modelling.html).

3.1 Topic identification

For each source/month pair, the average monthly term frequency was calculated by dividing the number of occurrences of the term by the number of occurrences of all terms.
The words have been transformed to their stemmed (base) form using SnowballStemmer. Afterwards, the weighted average of frequencies by source has been calculated. Weights have been assigned to ensure that no source has excessive influence on final results due to the number of articles and to maintain relative balance between American and other sources.
The weights are presented in the Appendix.

For all terms which occurred at least once in the last two months of the analysis, an ordinary least squares regression has been performed for the entire time period. The dependent variable of the estimation is the weighted frequency, while the number of months since the beginning of the analysed period is the independent variable. The result is a single coefficient (referred to as *coef*) β . Terms with the highest β coefficients have grown the most. However, the top growing words are always stopwords due to their sheer number of occurrences. Most lists of stopwords are not domain-specific: NLTK's list does not include words such as "internet", which should be regarded as a stopword in modern technological media. Instead of creating a domain-specific stopwords list, we divided the coefficient by the mean weighted frequency in all months of the regression. The resulting normalised coefficient (*coef_norm*) delivers a number which can be used to winnow out irrelevant terms by setting a threshold a term needs to achieve to be included in further analysis. The threshold has been set to 0.0125, a value high enough to remove stopwords (including domain-specific ones), but low enough to allow the capture of early signals of new technologies and quickly growing established topics.

In the case of the COVID-19 analysis, our methods have been adjusted for the significantly shorter time span (only six months). For this reason, we focus on changes in weekly (instead of monthly) frequencies. Similarly, we do not weigh sources in this analysis, but treat each article the same way. The rest of the regression analysis is conducted the same way (the considered terms occurred at least once in the last two weeks).

In both cases, the 1,000 terms growing most significantly (with the largest coefficient which are above the threshold for normalised coefficient) have been reviewed, and the relevant terms for further analysis have been selected.

3.2 Co-occurrence analysis

For the terms chosen in the previous part (“analysed terms"), the most common “co-occurring" terms out of the top 15,000 most significantly growing terms were calculated. The terms co-occur if they are present in the same article, not necessarily in the same sentence or paragraph. The number of occurrences of the co-occurring term in all articles in a given source containing the analysed term was checked. This number of co-occurrences was divided by the number of occurrences of the analysed term in all articles in the source. In the case of the first analysis, sources were aggregated using weights, just as frequencies were. The value of a term’s co-occurrence with itself is 100 when the term occurs in all sources.

3.3 Sentiment analysis

The same words which have been chosen for the co-occurrence analysis were selected for the sentiment analysis. For each term, the 100 most co-occurring terms were selected for further analysis. The sentiment analysis has been prepared using VADER [4], an open-source lexicon and rule-based sentiment analysis tool. VADER is specifically designed for social media analysis, but can be also applied for other text sources. The sentiment lexicon was compiled using various sources (other sentiment data sets, Twitter etc.) and was validated by human input. The advantage of VADER is that the rule-based engine includes word-order sensitive relations and degree modifiers. As VADER is more robust in the case of shorter social media texts, the analysed articles have been divided into paragraphs.

All paragraphs in the articles containing the analysed and co-occurring terms were modified to exclude these terms and assigned a score between -1 (most extreme negative) and 1 (most extreme positive) by VADER. The removal of the terms is necessary, as they may not be emotionally neutral, for example when some technologies or companies attempt to solve a negative issue. In such a case, the paragraph’s scores could be positive, but the negative terms would bring the paragraph's score down. Terms present in most negative and most positive (on average) paragraphs are then extracted.

[4] https://github.com/cjhutto/vaderSentiment

3.4 Topic modelling

Topic modelling assumes that each article is a mixture of topics, and each topic is a collection of characteristic terms. The most popular method of topic assignment is the Latent Dirichlet Allocation (LDA) model[5], which is a probabilistic topic model using Bayesian formulation to reveal hidden (latent) topics in the given text corpus. Documents in the corpus are treated as bag-of-words, i.e. the word ordering is not taken into account. The topics obtained via LDA analysis are probability distributions over terms. Each topic consists of a different set of terms characterised by a certain probability of appearance in the given subset of texts[6]. Instead of beginning the analysis with a predefined set of terms and codes derived from the domain expertise, the researcher specifies the number of topics the algorithm is supposed to find. The additional key hyper-parameters of the algorithm include α for the prior topic-document distribution and β for the prior word-topic distribution.

Designating a specific number of topics is considered to be a challenging process in text mining analyses[7]. LDA models are often evaluated based on the log-likelihood value for a held-out test set. However, Chang et al. (2009)[8] showed that predictive likelihood can differ significantly from the human assessment of the quality of topics.

In this analysis, we utilize a state-of-the-art coherence framework proposed by
Röder et al. (2015)[9]. The method assesses the coherence of topics by analysing the pairs of N most frequent topic-words. We optimize the coherence value with respect to the number of topics, α and β. Topic modelling was conducted using the Gensim package. Interactive visualizations of the obtained topics were prepared with the pyLDAvis package.

3.5 t-SNE

t-SNE is a dimensionality reduction technique: for our study, we would like to visualise articles in two dimensions based on their content.
SNE, the base algorithm for t-SNE, minimizes the difference between probability distributions for full-dimensional and low-dimensional embeddings. The goal is to preserve conditional probabilities that a point would choose another point as its neighbour. The distributions are Gaussian, and the standard deviation for a given point (it is not constant for all points) is chosen based on perplexity, which can be understood as the number of nearest neighbours considered. Gradient descent finds the local minimum.

---

[5] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, J. B. Edu, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003).
[6] D. M. Blei, Probabilistic topic models (article), Communications of the ACM (2012).
[7] G. Ignatow, R. Mihalcea, Text Mining: A Guidebook for the Social Sciences, 2018.
[8] J. Chang, J. Boyd-Graber, G. Sean, W. Chong, D. M. Blei, Reading Tea Leaves: How Humans Interpret Topic Models, Advances in neural information processing systems (2009).
[9] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, ACM, pp. 399–408. (2015)

As in the case of LDA, first our documents are transformed into bag-of-words matrix form. This matrix is high-dimensional: each row is a document and each column correspond to a different word and contains the number of its occurrences in the documents. Our task is to reduce the number of dimensions to two. Following the use of various techniques (SVD, PCA) that meant to reduce computational complexity and find initial structure, t-SNE is implemented. We will use this combination of methods to cluster articles in a two-dimensional space: articles covering similar topics will be placed close to each other by the algorithm.

4. Topic identification

4.1 Online news

The first step in our analysis is the identification of trending terms based on the changes in term frequencies. From the 1000 most trending terms, the NGI related terms are selected. The terms were assigned to wider thematic groups that are presented at Figures 3 (technologies) and 4 (social issues). The trending terms are also summarised as interactive visualisations on the website.

The figures present the most important technologies and social issues from the previous years. The emerging technologies include AI, 5G, quantum computing, decentralized computing and cryptocurrencies. The trending terms reveal the areas of implementation, such as AI - facial recognition, Google Duplex; cryptocurrencies - Libra, ico (initial coin offerings).

The identified social challenges all have important technological angles: the content crisis on social media, privacy, the state of democracy, competition in the digital economy, conflicts with China, online discrimination, climate change and the ethical use of technology.
During the analysis, we will focus on the relationship between tech and these highly relevant challenges.

Figure 3 Umbrella topics and identified keywords: Emerging technologies

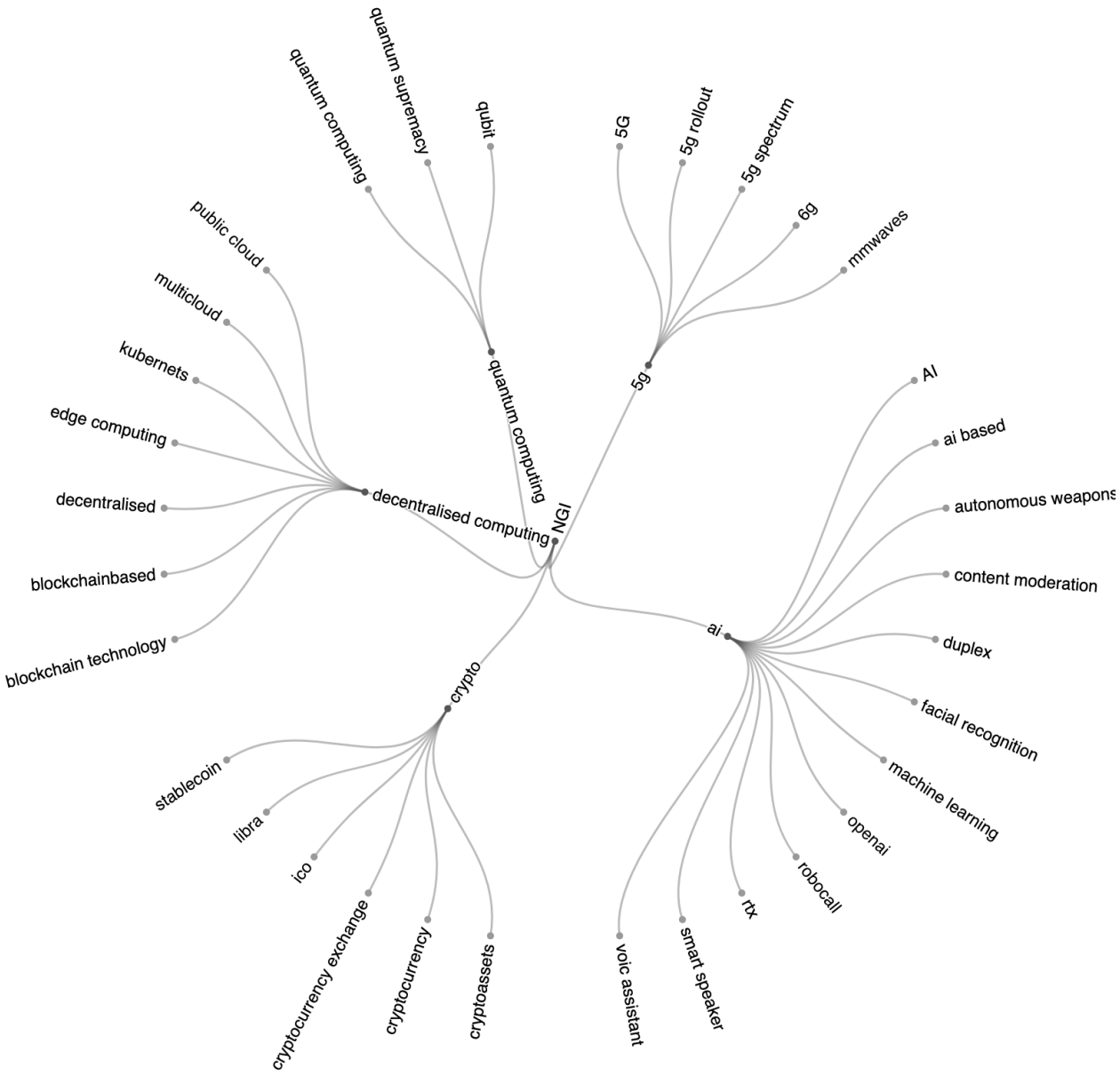
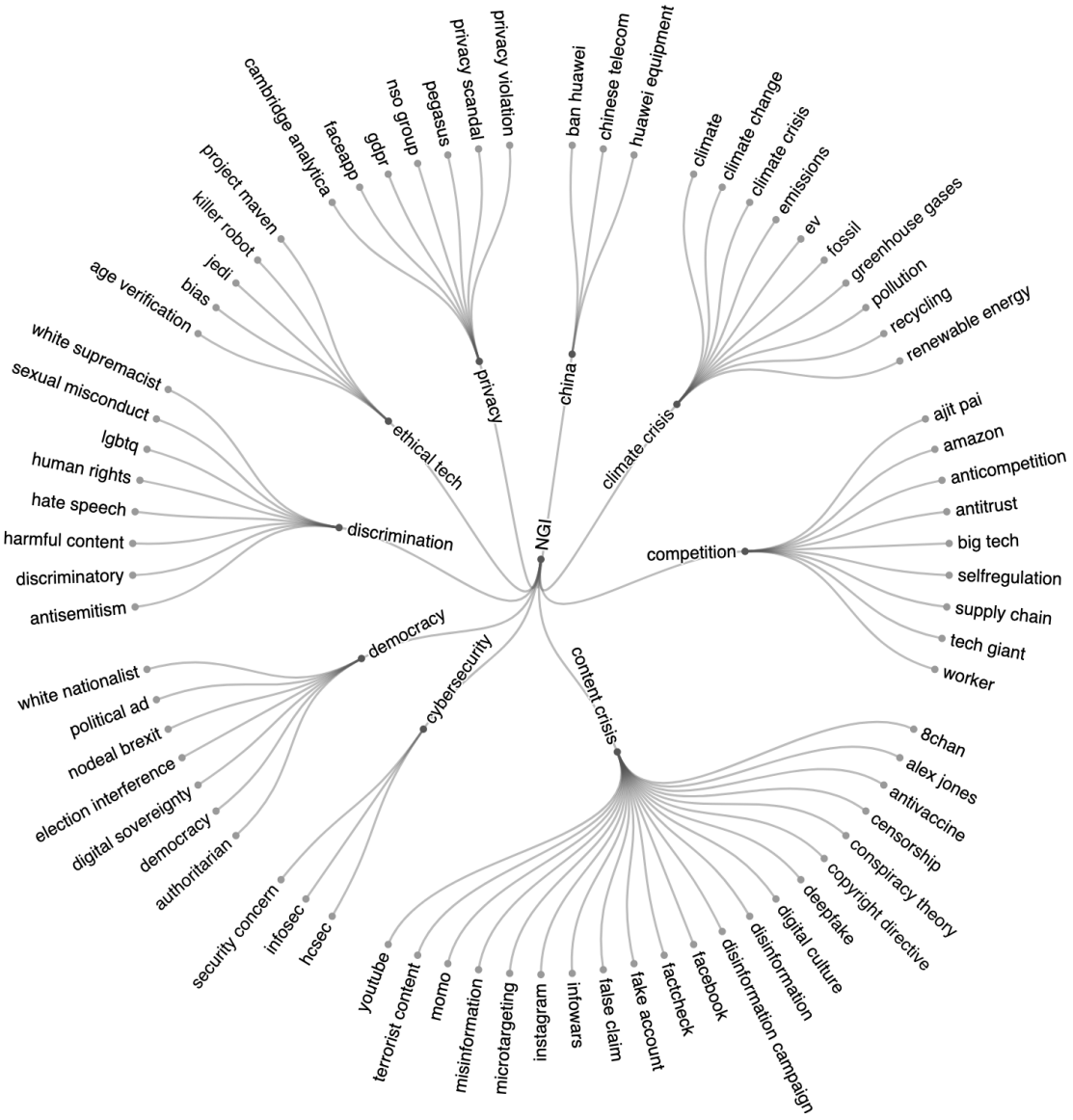


Figure 4 Umbrella topics and identified keywords: Relevant social challenges



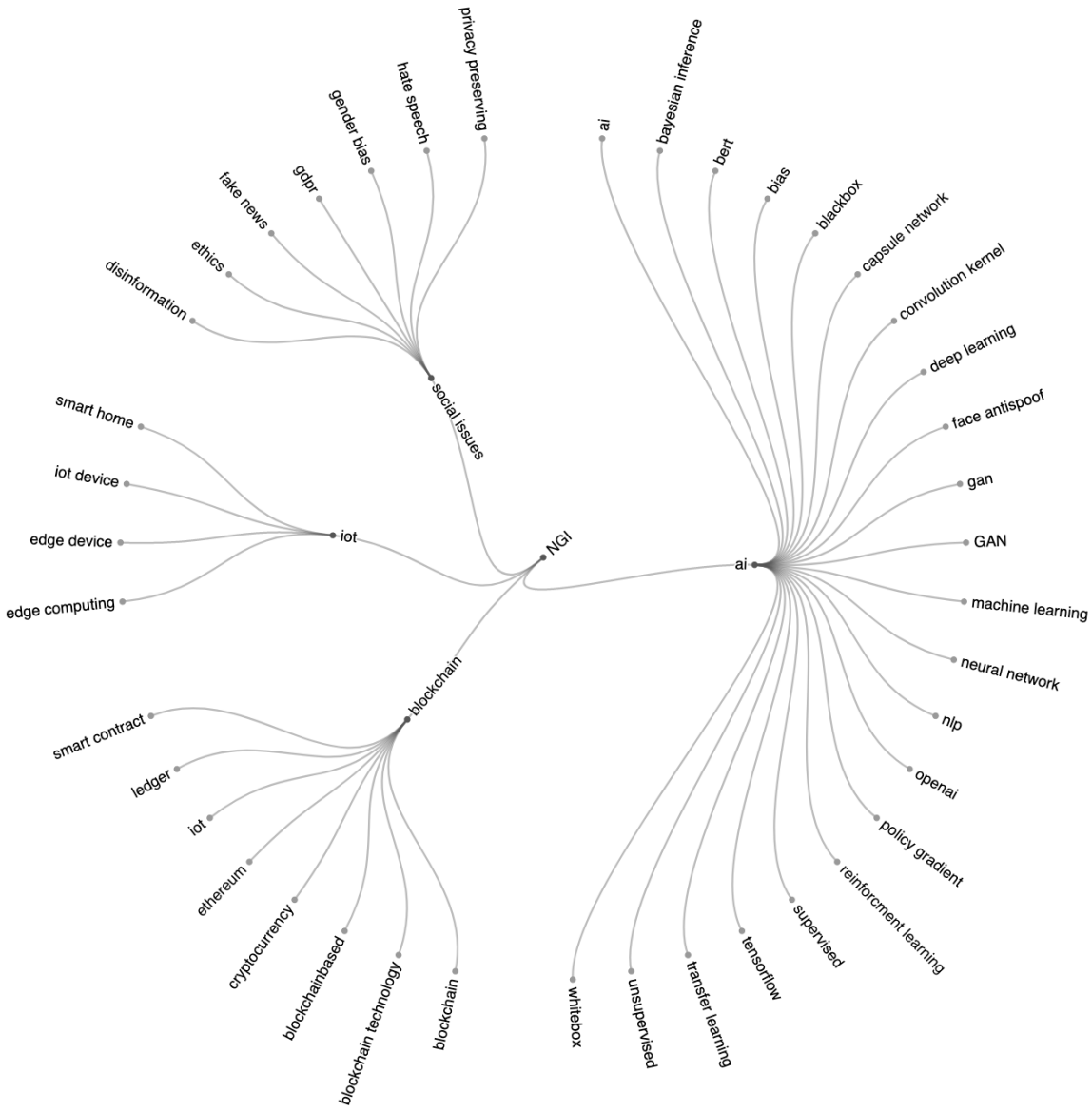
4.2 Working papers

Next, the abstracts of papers in arXiv and SSRN are analysed with the same method. Similarly to news articles, relevant terms are selected among the most trending terms.

Similarly to our previous analysis from June 2019, various AI and ML methods are trending in arXiv papers, including reinforcement learning (policy gradient) and generative adversarial learning (GAN). The trends also reveal methods related to image processing (convolution kernel) and natural language processing methods (Bert), as well as popular libraries and APIs (Tensorflow, OpenAI).

The results also support growing research interest in the field of decentralised technologies, such as blockchain or edge computing, and Internet of Things devices. Finally, while computer science and natural science researchers make up the main community of arXiv, several social challenges were identified in the papers, including fake news, online privacy or gender bias.

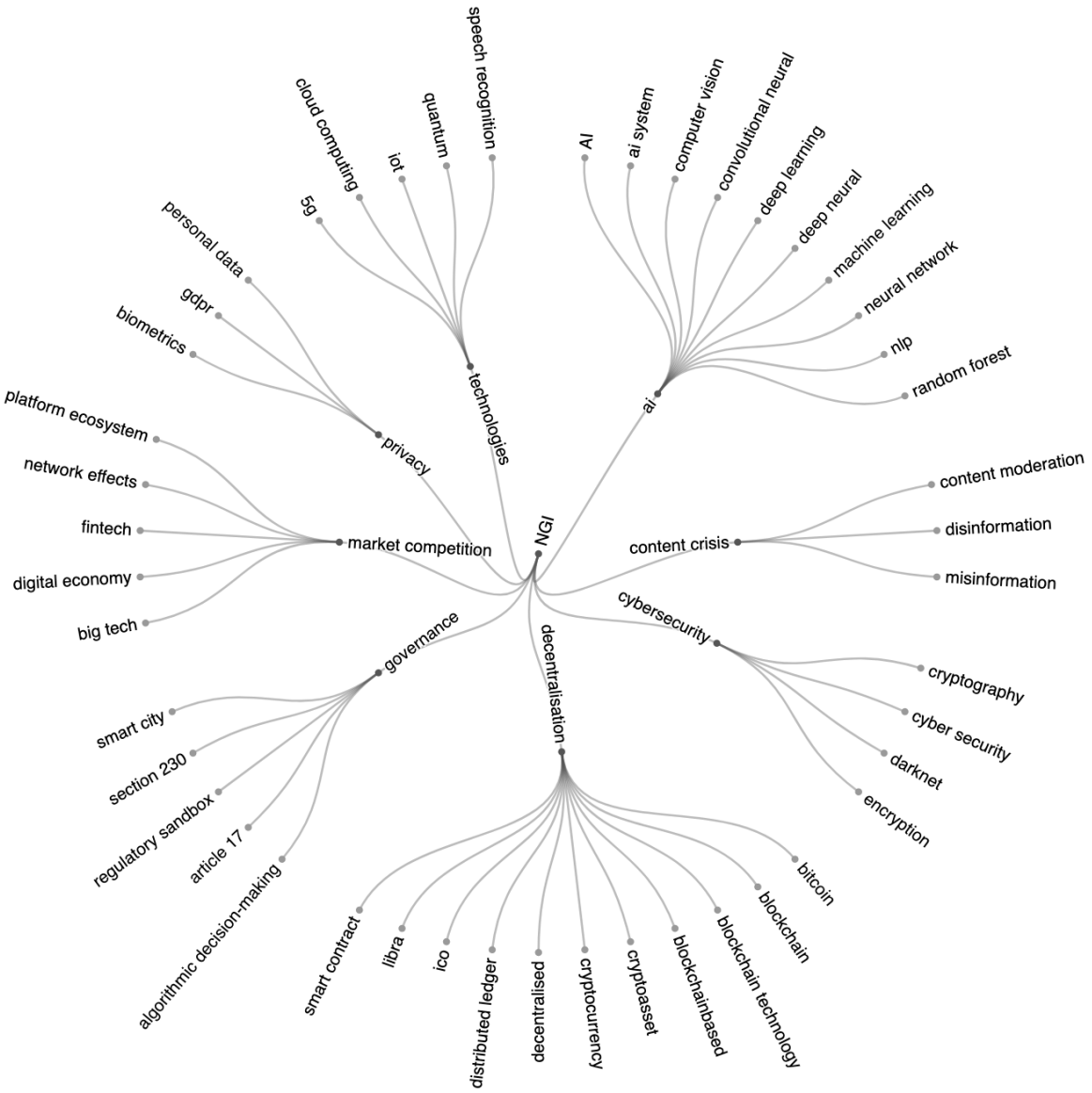
Figure 5 arXiv most trending terms



For better identification of trending areas in social science, the analysis is repeated on SSRN working paper abstracts. The most trending NGI related terms are summarised in Figure 6.

Interestingly, AI and ML methods have been trending in the social sciences as well. However, insights from arXiv provide more narrow and specialistic terms. Other common topics with arXiv include decentralised technologies. Researchers have increased their focus in the field of the content crisis (misinformation, content moderation), issues related to governance (algorithmic decision-making) and market competition (platform competition, digital economy).

Figure 6 SSRN most trending terms



5. Deep dives

The previous iteration of the topic identification process resulted in 8 key NGI topics (described in deliverable 1.9). These topics address wide concepts, combining various regulatory challenges and technologies:

- Trustworthy Information Flows: content crisis at social media and challenge for democracy
- Decentralised Power on the Internet: blockchain technologies, decentralised computing and competition policy
- Personal Data Control: privacy and cybersecurity
- Sustainable and Climate-friendly Internet: global warming
- Safer Online Environments: issues of hate speech and harmful content
- An Inclusive Internet: democracy, media literacy and 5G infrastructure
- Competitive European Ecosystems: emerging technologies and competition policy
- Ethical Internet Technology: the ethical use of AI

Naturally, the new lists of identified trending terms from Section 4 can be also arranged around these highly relevant topics. However, some of the topics contain too many different distinctive areas, such as the topic “Competitive European Ecosystems”, which discusses emerging technologies as well as competition and antitrust policy. In order to provide more focused case studies with the text-mining tools, we have selected more narrow areas. As an example, the issue of democracy or competition appeared in multiple topics, hence we will analyse them in greater detail separately.

In this report, we will focus on the following areas:

- Trustworthy Information
- Blockchain and cryptocurrencies
- Online Privacy
- Sustainability and Climate Crisis
- Safer Online Environments
- Democracy
- Market Competition
- Ethical AI

We will implement three text-mining methods for each topic: the analysis of co-occurrences, sentiment analysis and topic-modelling.

First, the co-occurrence analysis shows which terms are related: as an example, which are the frequent terms in articles including the word “climate change”. Second, we also compute the sentiment score for paragraphs including frequent word pairs (such as climate change) to establish if a co-occurrence is mentioned in a more positive or negative context. This helps us to identify the relationship between concepts, institutions, and actors. These two steps are based on the expert analysis of results: the presented terms are selected from a list of most

relevant terms (e.g. in the case of co-occurrences, the top 20 most frequently appearing terms).

Finally, we explore the topics discussed in news articles that contain specific trending terms (“climate change”, “co2 emission” etc) using a popular unsupervised machine learning method - LDA. The list of terms used to select the articles is presented in the Appendix. This method provides less expert input, hence it is a useful robustness check for our analysis.

5.1 Trustworthy information

Digital disinformation, fake news and conspiracy theories are increasingly threatening democratic societies. The combination of various methods provides rich details on the news coverage on these issues. First, the co-occurrence analysis marks the challenges posed by the surging number of bots spreading fake news all over the social media and AI methods like “deepfakes”, making it simpler to doctor video and audio (Figure 7).

Second, the sentiment analysis helps to establish the positive and negative relationship between the terms (Table 1, Table 2). The most negative terms reveal a relentless interest about Alex Jones - one of the most infamous conspiracy theory spreaders who was banned from most of the online platforms. Donald Trump is also mentioned in this context, after being accused by fact-checkers of spreading misinformation, dividing public opinion over freedom of speech and online censorship. This recently led to Facebook and Twitter blocking a video shared by accounts linked to the president for violating their policies on coronavirus misinformation.

The topic modelling exercise indicated the important role AI research plays in the disinformation debate both as a threat and potential game-changer thanks to next-generation content moderation (Table 3). Other major topics were related to election meddling via online platforms and disinformation related to 5G, climate crisis and Wikileaks.

Figure 7 Co-occurrence analysis for Trustworthy Information

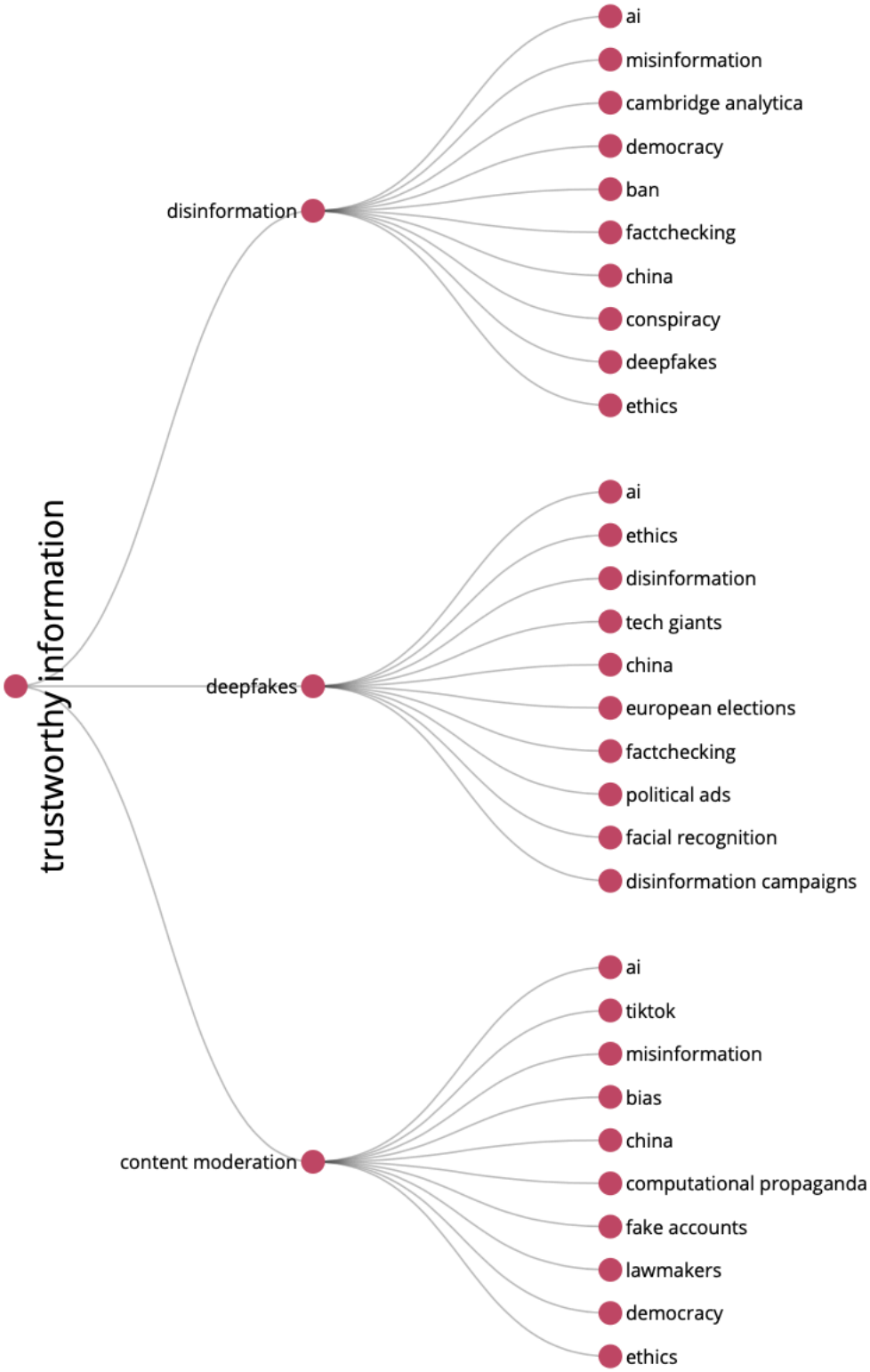


Table 1 Co-occurrences with most positive and negative sentiments: **content moderation**

Most positive	Most negative
trustworthy recognition gdpr midterm election warren	nationalist graphic violence hateful content self harm antisemitism

Table 2 Co-occurrences with most positive and negative sentiments: **disinformation**

Most positive	Most negative
cryptocurrency european election 5g libra gdpr	espionage infowars alex jones president trump conspiracy theory

Table 3 Topic modelling: Trustworthy Information

Topic 1 AI (53.7% of tokens)	Topic 2 Online platforms (26.6% of tokens)	Topic 3 Smart devices (8.8% of tokens)
ai startup business technology game data app microsoft amazon research	facebook twitter youtube content policy google platform social media election instagram	apple device game android camera app iphone smartphone google hardware
Topic 4 5G and cloud (4.4% of tokens)	**Topic 6 Climate crisis (2.5% of tokens)**	**Topic 7 Wikileaks (0.6% of tokens)**
5g cloud kubernets vmware aws security huawei infrastructure 5g network public cloud	pichai musk space x nasa climate change moon boe mars launch storm	assange charges indictment extradite fraud prosecutor conspiracy ddos attack wire fraud hack

5.2 Blockchain and cryptocurrencies

In the last few years, we have observed blockchain and cryptocurrencies emerging into the mainstream. However, we still do not see problems that are entirely solved by these technologies. We often hear that “Blockchain is a solution looking for a problem”[10].

The most frequent co-occurrences for cryptocurrencies include Libra, a private blockchain digital currency proposed by Facebook and stablecoin, a class of cryptocurrencies attempting to offer price stability and being backed by a reserve asset. Stablecoin is also present among most positive co-occurring terms with Libra and Calibra. Calibra is a Facebook subsidiary whose goal is to support the adoption of Libra with various services, or as summerised by the Verge: ‘Facebook’s secret weapon for monetizing its new cryptocurrency’[11]. Hyperledger Fabric - an open-source, permissioned distributed ledger developed by the Linux Foundation also received a positive reaction in the tech media.

The most negative stories include terms related to Venezuelan cryptocoin Petro. Petro’s value depends on the national oil reserves and is the first national cryptocurrency. The cryptocoin was created to support the plummeting economy of Venezuela and was banned by some countries, including the USA.

The topic modelling exercise showed the media interest in the challenges blockchain poses to the existing legal and regulatory framework, its potential impact on other emerging technologies, and cybersecurity issues related to decentralised ledgers.

[10] https://www.forbes.com/sites/darrynpollock/2020/01/29/will-2020-be-the-year-cryptocurrency-and-blockchain-becomes-operational/

[11] https://www.theverge.com/2019/6/18/18682838/facebook-digital-wallet-calibra-libra-cryptocurrency-kevin-weil-david-marcus-interview

Figure 8 Co-occurrence analysis for Blockchain and crypto

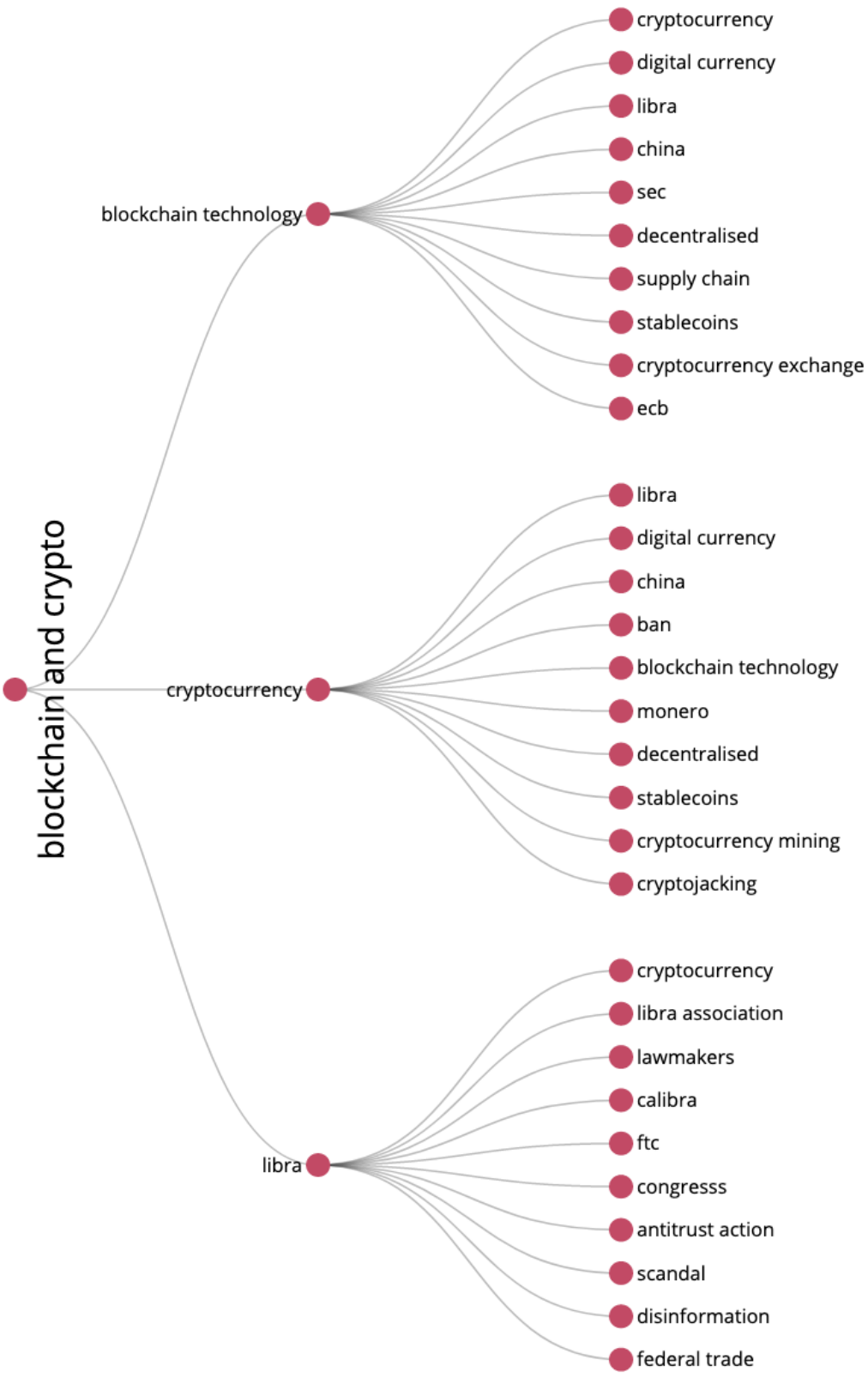


Table 4 Co-occurrences with most positive and negative sentiments: **blockchain technology**

Most positive	Most negative
hyperledger fabric supply chain decentralisation consent digital identity	venezuela shutdown ecb protests sanctions

Table 5 Co-occurrences with most positive and negative sentiments: **libra**

Most positive	Most negative
coinbased fiat currency stablecoin calibra central banker	political ad political party misinformation president donald ftc

Table 6 Topic modelling: Blockchain and crypto

Topic 1 Cybersecurity (51.2% of tokens)	Topic 2 Regulations (29.4% of tokens)	Topic 3 Online services (12.3% of tokens)
facebook malware attack user startup app data blockchain research team	facebook blockchain cryptocurrency regulation bitcoin bank libra payment trade government	app apple google disrupt berlin android game spotify streaming disney microsoft
Topic 4 Impact on other technologies (5.4% of tokens)	Topic 6 Finance (0.3% of tokens)	Topic 7 Cryptocurrencies (0.2% of tokens)
cloud aws microsoft huawei nvidia iot intel ai	xrp atari softbank fsb yuan ripple financial stability bitmain	coinbased check account robinhood telegram europol bitcoin chain analysis bug bounty

chip 5g	venezuela iceland	cryptocurrency exchange ethereum classic

5.3 Online privacy

Much has changed for online privacy in recent years. GDPR was a landmark for personal data protection in the EU and beyond, and the debate is now largely about how to regulate online platforms, not whether regulation is required.

Our societies have acknowledged that human rights are abused at population scale because of pervasive profiling, which brings us closer to the reign of surveillance capitalism. We observe recurring privacy concerns over new social apps fads like Faceapp, the AI powered selfie-editing app which uploaded users' photos to the cloud without making it clear to them that processing is not going on locally on their device[12], or Chinese TikTok, which was accused of failing to protect children's privacy by the US FTC.

The co-occurrence analysis points to the controversies about Israeli spyware company NSO Group and its main product Pegasus - malware able to intercept communications sent to and from a device, including communications over encryption-using apps like iMessage or Telegram. Sentiment analysis brings to light the most controversial GDPR violations including Marriott's data breach and TikTok being under investigation over child data use. The positive stories include issues related to the California Consumer Protection Act which came into effect in the beginning of 2020 as well as the principle of data portability being guaranteed in the EU by GDPR, and privacy promoting companies like Mozilla.

Topic modelling provides a similar picture. Tech media reported mostly on tech giants' and smart devices' role in the privacy ecosystem, novel privacy regulations, links between privacy and cybersecurity issues and campaigns against tech giants abuse of personal data.

[12] https://techcrunch.com/2019/07/17/faceapp-responds-to-privacy-concerns/

Figure 9 Co-occurrence analysis for Online privacy

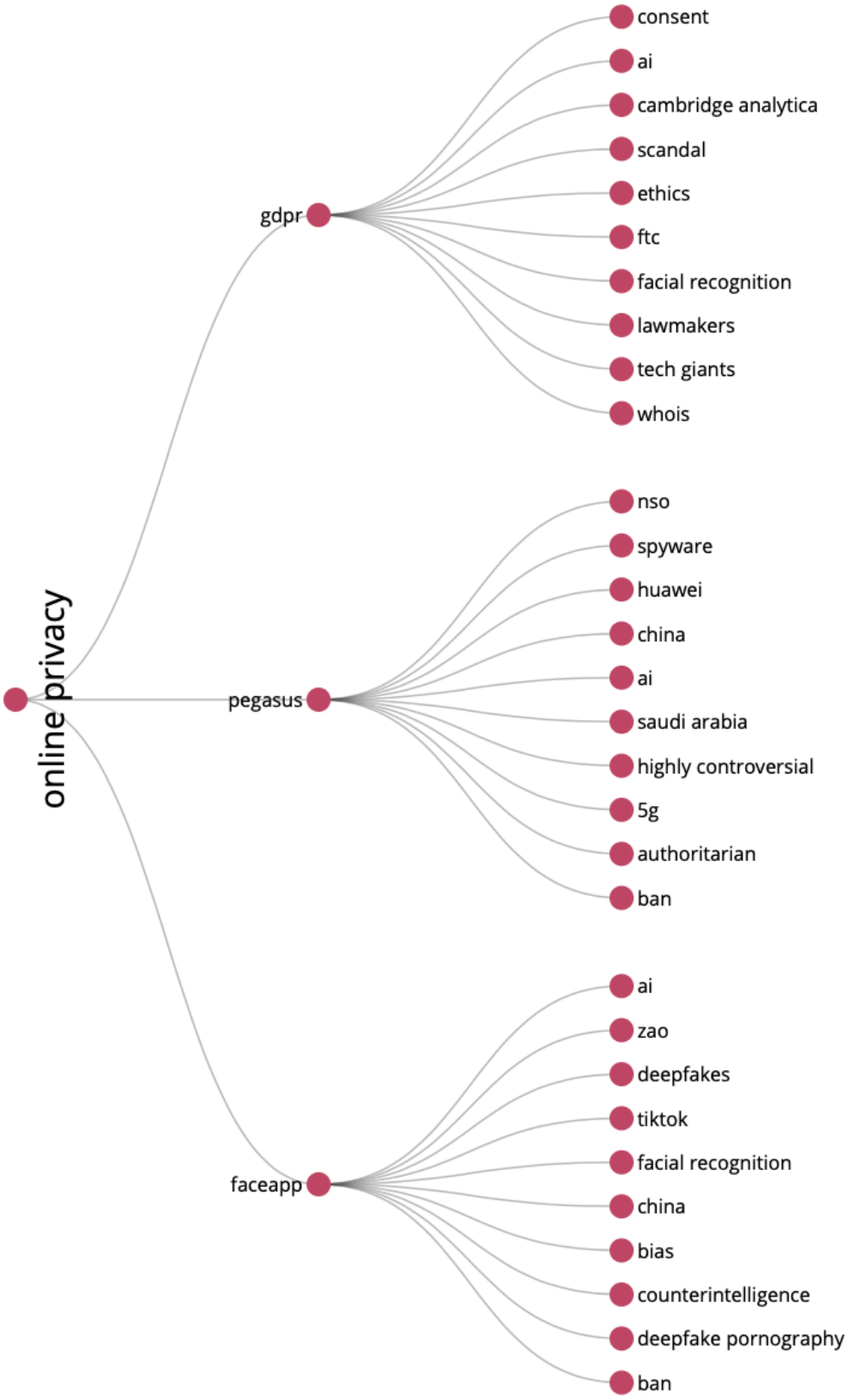


Table 7 Co-occurrences with most positive and negative sentiments: **faceapp**

Most positive	Most negative
ai facial recognition ethic ftc consent	goncharov bias tiktok chinese ban

Table 8 Co-occurrences with most positive and negative sentiments: **gdpr**

Most positive	Most negative
ccpa voice assistant face recognition data portability mozilla	whois marriott tiktok disinformation irish data

Table 9 Topic modelling: Online privacy

Topic 1 Tech giants (63% of tokens)	Topic 2 Security (21.5% of tokens)	Topic 3 Smart devices (5% of tokens)
facebook google apple amazon ai platform ads facial recognition microsoft instagram	zuckerberg government security blockchain policy breach law enforcement bank law attack	android camera apple huawei samsung iphone vulnerability smart home security bug
Topic 4 Controversies (3.2% of tokens)	**Topic 5 Content moderation (1.8% of tokens)**	**Topic 8 Regulations (1.1% of tokens)**
voice assistant brave pegasus 5 billion 2fa denham microtargeting location services musk fitbit	zuckerberg content moderation icann tencent monopoly fake news wechat duckduckgo sexual harassment conspiracy theory	age verification dpc schrem irish data cjeu high court sccs ukip justice ministry ecj

5.4 Sustainability and Climate Crisis

The co-occurrence analysis shows the most important trending terms related to technologies (electric vehicles), policy proposals (Green New Deal in the US) and companies (Dell, car manufacturers). This provides us with an overview of the main actors and recent news stories (e.g. companies committing to reduce their carbon footprint[13]). The positive news stories are related to promising technologies: heat pumps used in buildings[14], electric vehicles, progress in the field of renewable energy and food industry[15], computations made possible by AI[16] and quantum computing. On the other hand, news with negative sentiment reported on the effects of global warming (wildfires, heat waves), and also about the social unrest surrounding this challenge. A further pressing issue is related to the problem of fake news, conspiracy theories and science scepticism.

The topic modelling analysis also provides similar categories of discussions. Based on the frequent topic words, tech news reported mostly on EVs (Topic 1), tech giants (Topic 2) and policy issues (Topic 3). Topic 5 provides a list of influential actors, including politicians, institutions (US Environmental Protection Agency), companies and tech leaders.

Climate change remains humanity's top challenge, with great impact on technological and social development. Besides already available consumer products (EVs), emerging technologies such as AI and quantum computing may play a significant role in reducing the harmful effects of global warming. However, the content crisis on social media divides society by popularising fake news. Therefore, Internet services play a greater role in the fight against climate change that is beyond the carbon footprint of using them. Reducing the spread of fake news and propaganda will be key to build a global consensus in the necessity to take more significant steps.

[13] https://www.theverge.com/2019/11/12/20960357/dell-greener-computers-recycling-climate-change-pledge
[14] https://spectrum.ieee.org/energywise/energy/environment/heat-pumps-could-shrink-the-carbon-footprint-of-buildings
[15] https://www.theguardian.com/environment/2019/jan/16/new-plant-focused-diet-would-transform-planets-future-say-scientists
[16] https://www.theverge.com/2019/6/25/18744034/ai-artificial-intelligence-ml-climate-change-fight-tackle

Figure 10 Co-occurrence analysis for Sustainability and Climate Crisis

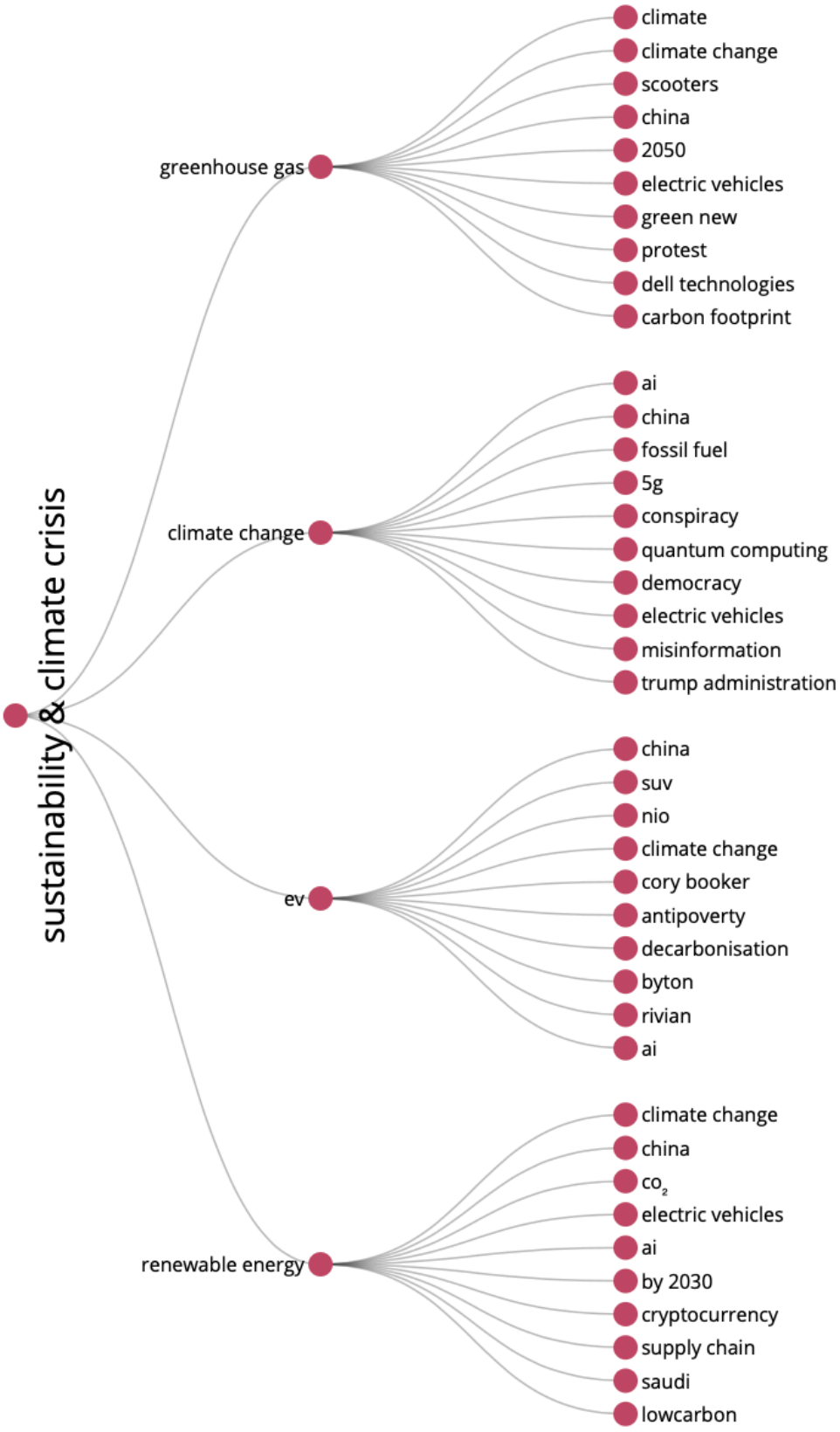


Table 10 Co-occurrences with most positive and negative sentiments: **greenhouse gas**

Most positive	Most negative
heat pump autonomous plant-based solar wind electric vehicle	geoengineering climate emergence extinction protest democracy

Table 11 Co-occurrences with most positive and negative sentiments: **climate change**

Most positive	Most negative
electric vehicle ai Bezos qubit supply chain	deep fake conspiracy wildfir vaccin heatwaves

Table 12 Topic modelling**:** Sustainability and Climate Crisis

Topic 1 EV (74.8% of tokens)	Topic 2 Tech giants (11.8% of tokens)	Topic 3 Climate crisis (6.4% of tokens)
amazon google car ai data battery vehicle tesla emission electric	Apple Facebook 5G Huawei Samsung Uber Google Volkswagen China investor	climate crisis EU storm ice commission degree celsius record united nations glacier heat wave
Topic 4 Fires (2.6% of tokens)	**Topic 5 Actors (1.2% of tokens)**	**Topic 7 Energy (0.8% of tokens)**
fire species extinction meat forest animal blaze soil fire season temperature	Bezos EPA Trump Merkel climate talk tariff Walmart Paris Agreement Vestager Pichai	coal Trump administration natural gas reef oil CO2 offshore wind gas fossil fuel plant

5.5 Safer online environments

Making the Internet a safer place is one the greatest digital challenges for governmental and non-governmental actors. Big tech increasingly uses algorithms to combat the spread of discrimination, hate speech, online extremism and child exploitation. In 2019 alone, Facebook claims to have flagged almost 20M hate speech posts using AI systems before they were reported by users[17]. The last evaluation of the EU Code of Conduct shows that companies are now assessing 90% of flagged content within 24 hours and 71% of the content deemed illegal hate speech is removed[18].

Despite significant developments in automatic hate speech recognition, much controversy about content removal remains. In the debate referring to human rights narrative we observe tensions between freedom of speech and the right to be protected from any form of discrimination.

Sentiment analysis reveals important actors in the online discrimination debate, e.g. Jonathan Zittrain - professor of Internet law as well as people and websites accused of hate speech violations, e.g. Tommy Robinson or 8chan.

Additionally, topic modelling analysis shows that politically motivated online extremist content is becoming a more pressing issue as online outlets promoting white supremacy and antisemitism are gaining popularity (e.g. Daily Stormer).

[17] https://www.wired.com/story/facebook-ai-hate-speech-improves-unclear/

[18] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Figure 11 Co-occurrence analysis for Safer online environments

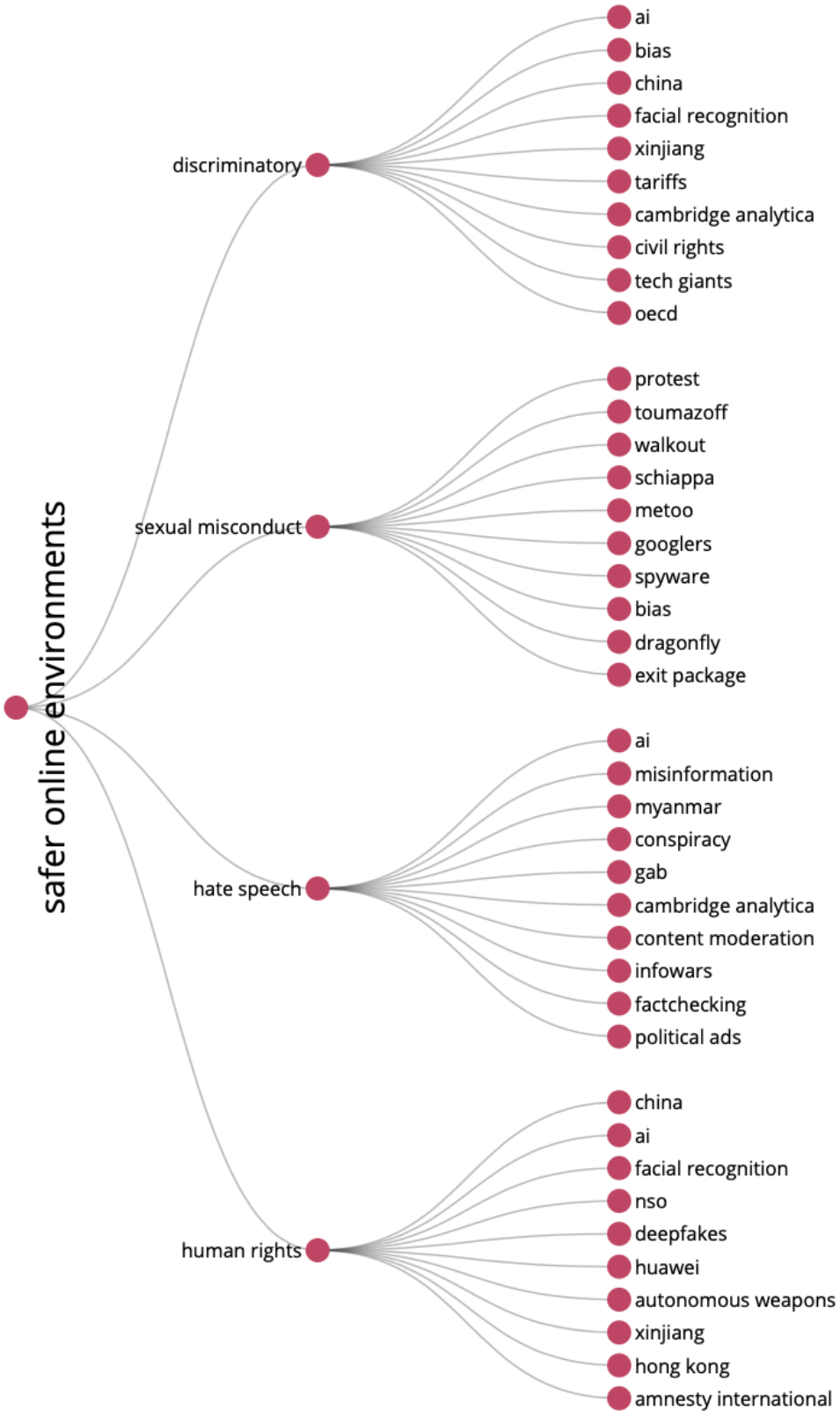


Table 13 Co-occurrences with most positive and negative sentiments: **hate speech**

Most positive	Most negative
jonathan zittrain gdpr berners lee consent factcheck	maza homophobic robinson rohingya 8chan

Table 14 Co-occurrences with most positive and negative sentiments: **human rights**

Most positive	Most negative
yuval noah supply chain gdpr 5g tech giant	weapon system rohingya nso group pegasus khashoggi

Table 15 Topic modelling: Safer online environments

Topic 1 Tech giants (66% of tokens)	Topic 2 Social media (12.4% of tokens)	Topic 5 European policy (2.1% of tokens)
facebook google app data amazon ai youtube platform ads instagram	facebook twitter trump ban policy china election social media violation youtube	brexit fcc legislation data protection tax human rights net neutrality european parliament european commission isp
Topic 7 Fake news (0.9% of tokens)	**Topic 8 Hate speech (0.8% of tokens)**	**Topic 9 Extremist content (0.8% of tokens)**
5g fake news vestager factcheck citizen lab whistleblower bezos trump administration indian government cobalt	fake account hate speech take down terrorist content illegal content jourova breitbart ansip right wing remove	8chan daily stormer grindr altright farright 2fa extremist content terrorist propaganda remove illegal hateful conduct

5.6 Democracy

The state of public debate is heavily influenced by social media, the spread of fake news and conspiracy theories. Since the Russian interference scandal in the US elections, and the Cambridge Analytica campaign during the Brexit referendum, there are increasing warning signs that elections can be manipulated through social media.

The most frequent co-occurrences for democracy include terms reflecting these issues: disinformation, political advertisements, tech giants, and the use of AI to manipulate facts (e.g. deep fakes). The most positive word-pairs include the fight against climate change, Tim Berners-Lee (founder of the word wide web, advocating for a human-centric Internet) and GDPR. On the other end of the spectrum, the most negative ones signal the issue of fake news and conspiracy theories (e.g. the chaos in Myanmar fuelled by Facebook[19]), the Chinese government and the far-right.

Among the threats for just and fair elections are targeted political advertisements. Various companies (such as Cambridge Analytica or Aggregate IQ) specialise in political campaigns on social media, spreading false information for targeted users. Moreover, these campaigns often breach campaign spending laws[20]. The co-occurrences show the related problems: computational propaganda ("the use of information and communication technologies to manipulate perceptions, affect cognition, and influence behaviour"[21]), the possible involvement of foreign agents and the spread of micro-targeted messages. Among the most negative relationships is the Mueller report[22] investigating Russian interference in the 2016 elections, and ads targeting the LGBTQ+ community[23].

The characteristic terms for news articles also reveal other issues, such as the issue of privacy, the political crisis in Hong Kong and the cybersecurity side of elections.

[19] https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/
[20] https://techcrunch.com/2018/07/26/facebook-finally-hands-over-leave-campaign-brexit-ads/
[21] https://www.politico.com/magazine/story/2018/02/04/trump-twitter-russians-release-the-memo-216935
[22] https://www.justice.gov/storage/report.pdf
[23] https://www.theguardian.com/technology/2019/dec/30/facebook-prep-hiv-ads-removal

Figure 12 Co-occurrence analysis for Democracy

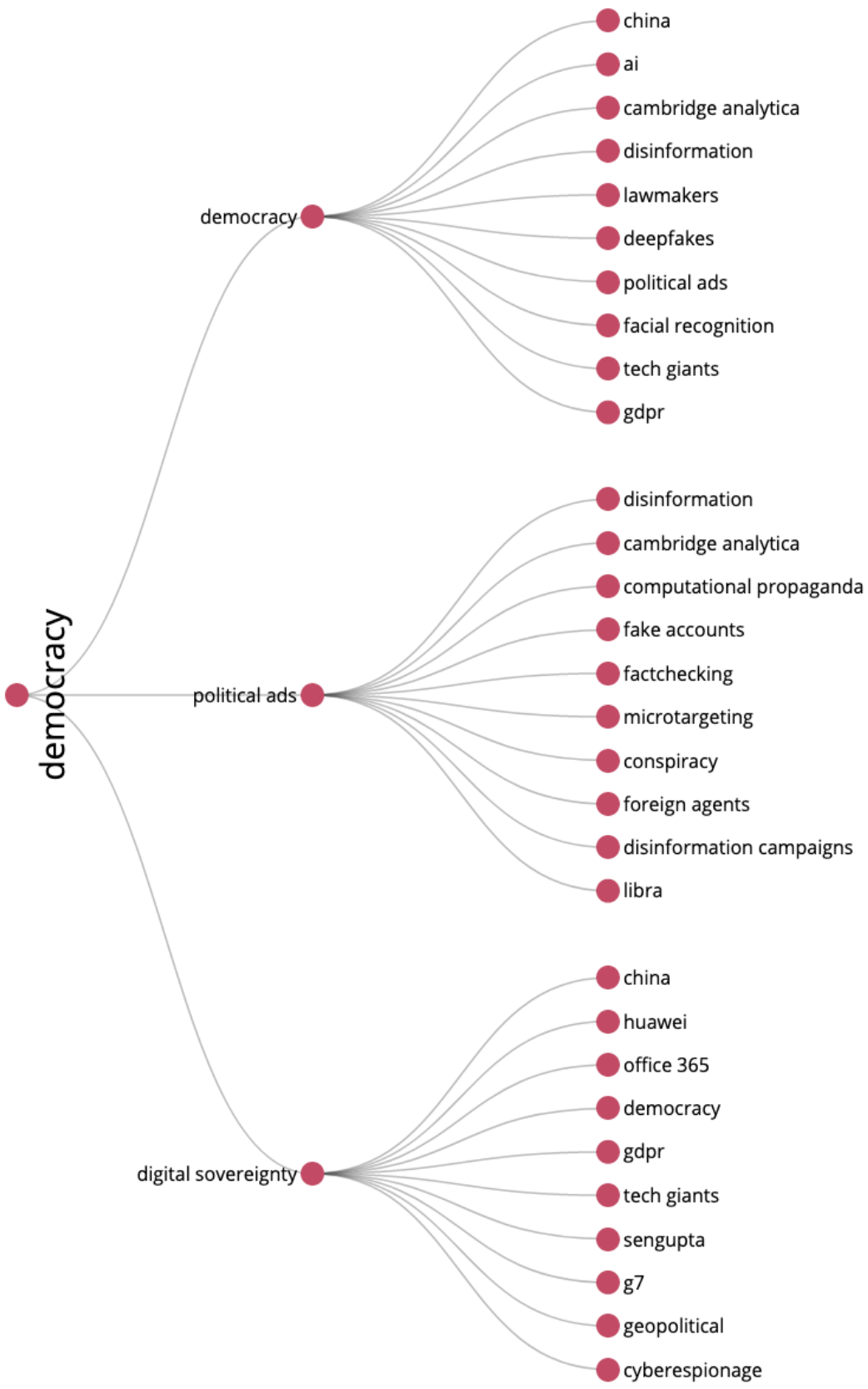


Table 16 Co-occurrences with most positive and negative sentiments: **democracy**

Most positive	Most negative
climate change migration berners lee gdpr big tech	myanmar farright free express conspiracy theory chinese government

Table 17 Co-occurrences with most positive and negative sentiments: **political ad**

Most positive	Most negative
media literacy factual information coordinated disinformation democratic process election interference	aiq authoritarian regime conspiracy theory lgbtq mueller

Table 18 Topic modelling**: democracy**

Topic 1 Political ads (50.9 % of tokens)	Topic 2 Privacy (18.7% of tokens)	Topic 3 Social media (17% of tokens)
political ad twitter google ban advertisement China regulation social media campaign disinformation	Google AI Dorsey Youtube Whatsapp Amazon privacy surveillance zuckerberg encryption	Trump Twitter lie political party policy social media Snowden campaign content ban
Topic 4 Tech and state (2% of tokens)	**Topic 5 China (1.4% of tokens)**	**Topic 8 Elections (0.8% of tokens)**
Huawei 5G facial recognition 2020 elections AI state law enforcement biometric scheme supreme court	Mozilla Wechat shutdown citizenship hong kong Vestager bug source code Chinese Ali baba	Harbath Cybersecurity paper ballot vote machine blockchain hack russia national security paper trail system

5.7 Market Competition

The digital transformation has a profound impact on the economy. Platformisation changed the market dynamics, facilitating the development of giant companies. Competition policy became highly relevant not only in the case of existing services (e.g. social media), but also for emerging technologies, such as cryptocurrencies or 5G.

The co-occurrence analysis provides the main axes of recent events: the “trade wars” between US and China, controversies surrounding tech giants and tech industries that lack proper regulation.

The terms related to “Chinese telecom” describe well the key moments of the trade wars: the ban of US companies to export products to Chinese ZTE in 2018[24], the extradition controversy of Huawei CFO Meng Wanzhou[25] (who has been accused to violate sanctions put on Iran) and the accusation of espionage and ban for US companies to conduct business with Huawei[26].

The technology areas that are in the centre of regulatory debate include cryptocurrencies (such as Facebook Libra), targeted political ads (especially since the Cambridge Analytica and S.C.L controversy[27]) and games (especially loot boxes that may lead to gambling problems[28]).

[24] https://www.theverge.com/2018/5/9/17336454/zte-us-phone-ban
[25] https://www.theguardian.com/technology/2019/may/09/meng-wanzhou-huawei-cfo-seeks-halt-to-extradition-after-trump-comments
[26] https://www.theverge.com/2019/3/17/18264283/huawei-security-threat-experts-china-spying-5g
[27] https://www.zdnet.com/article/how-cambridge-analytica-used-your-facebook-data-to-help-elect-trump/
[28] https://www.wired.com/story/loot-boxes-parliament/

Figure 13 Co-occurrence analysis for Market Competition

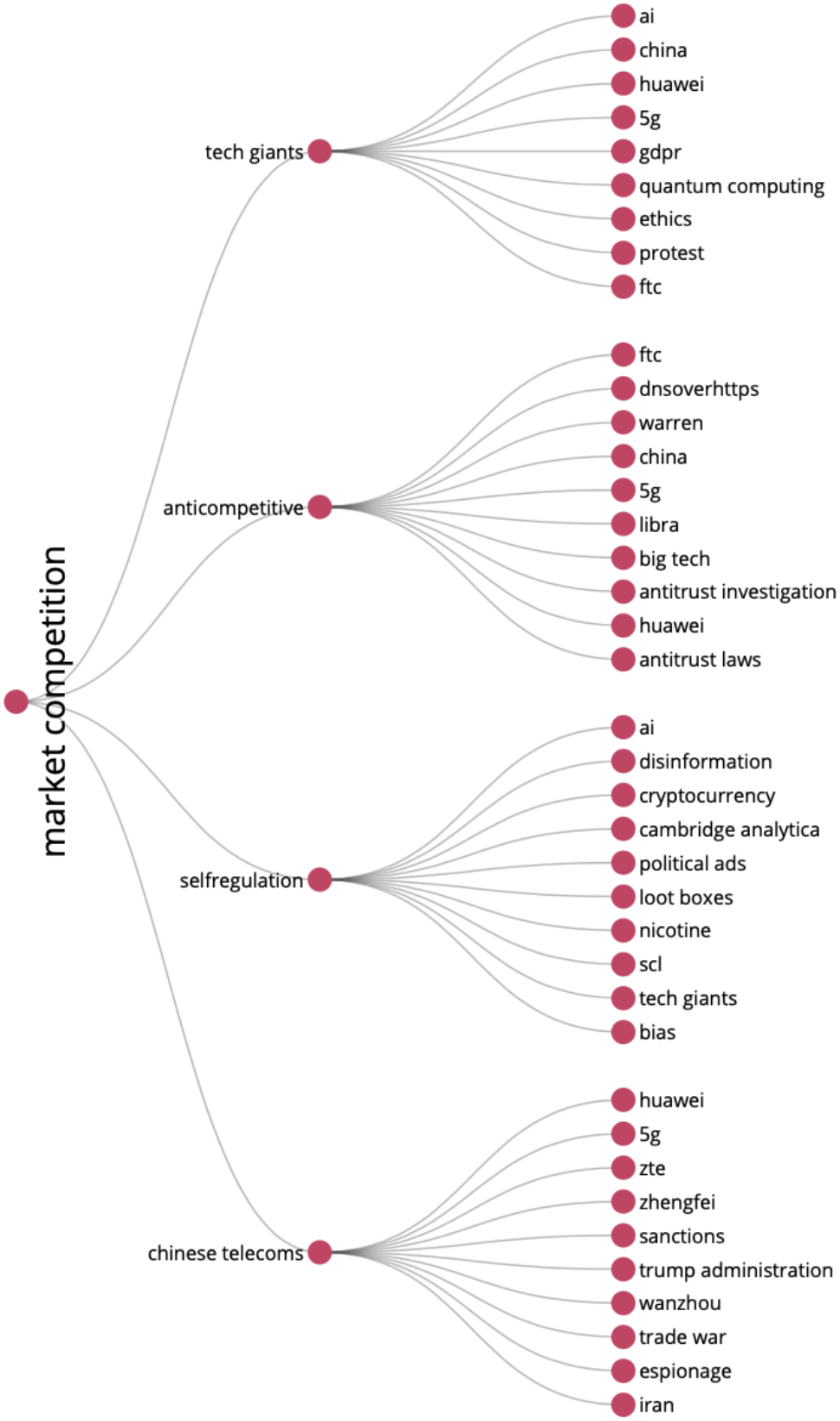


Table 19 Co-occurrences with most positive and negative sentiments: **tech giant**

Most positive	Most negative
ikea sono qubit smart speaker decentralised	infowars conspiracy tvcs content moderation political ad

Table 20 Co-occurrences with most positive and negative sentiments: **anticompetitive**

Most positive	Most negative
Libra association duckduckgo cryptocurrency section 230 fortnite	modem chip discriminatory €2.4bn protest clegg

Table 21 Topic modelling: market competition

Topic 1 Tech giants (42.5 % of tokens)	Topic 2 Data (38.9% of tokens)	Topic 3 China / chips (5.7% of tokens)
huawei google facebook amazon eu apple tax data ftc commission	Facebook app apple data google user AI platform instagram privacy	Huawei Qualcomm China Chip Intel Alibaba ZTE Apple Tencent AW
Topic 4 5G / Telco (2.6% of tokens)	**Topic 5 Content (1.2% of tokens)**	**Topic 7 Trade war (1.1% of tokens)**
Huawei 5G Vodafone ZTE network patent state attorney Cisco China mobile executive order	Alexa Google Assistant Disney Netflix Verizon Comcast smart home Youtube Siri Shazam	Qualcomm Broadcomm tariff trade war modem patent telecom equipment constant currency china Huawei chip

5.8 Ethical AI

We identified various technologies that raise crucial ethical questions. The fast development of artificial intelligence algorithms and their increased use in facial recognition and autonomous weapons are among the most important identified trends.

The use of facial recognition by companies, governments, law enforcement and the military has been in the centre of heated debates. Our analysis provides various examples for the risks and associated controversies.

The co-occurrences show that consumer products with facial recognition tech include smartphones and smart surveillance cameras. The latter product category has been frequently related to privacy concerns, such as in the case of Amazon Ring. Amazon has been reportedly developing facial recognition products and provided footage recorded with its surveillance camera for law enforcement[29]. The use of automated facial recognition (AFR) by police is especially controversial, as marked by the co-occurring terms with negative sentiments. The news stories include the debate on the use of AFR in England and Wales, the use of the technology to control the Uighur community in China[30], and the blacklisting of Chinese surveillance tech giant Hikvision[31].

The involvement of tech giants in military projects has been also widely discussed. A controversial example is project Maven, led by the US Department of Defence (DoD). Google cooperated in the project to use AI for drone image processing[32]. Following external backlash and the protest of Google employees, the company pulled out from the project and did not bid in other Pentagon projects (such as JEDI - Joint Enterprise Defense Infrastructure). On the other hand, the co-occurrences of project Maven with positive sentiment include OpenAI, a project aiming at developing responsible AI solutions[33].

The third example for the controversial use of technology is robocalls. Automated calls have been especially problematic in the US, where the FCC waged a regulatory war against them. The co-occurrences show various related technologies: the STIR/SHAKEN protocol[34] that can be used to authenticate calls, and Duplex, Google's released software for automated calls[35].

The topic modelling exercise revealed other areas as well: discussions on alleged bias and censorship in social media and the use of facial recognition at airports.

[29] https://www.theverge.com/2019/7/29/20746156/amazons-ring-law-enforcement-partnerships
[30] https://www.theguardian.com/technology/2019/jun/07/facial-recognition-technology-liberty-says-england-wales-police-use-should-be-banned
[31] https://www.zdnet.com/article/us-reportedly-to-blacklist-chinese-surveillance-camera-giant-hikvision/
[32] https://www.theguardian.com/technology/2018/mar/07/google-ai-us-department-of-defense-military-drone-project-maven-tensorflow
[33] https://openai.com
[34] https://www.theverge.com/2019/5/3/18525784/robocalls-spam-calls-solutions-comcast-at-t-t-mobile-stir-shaken-scam-john-legere-telecom
[35] https://arstechnica.com/gadgets/2019/04/google-duplex-arrives-on-iphones-most-android-devices/

Figure 14 Co-occurrence analysis for Ethical AI

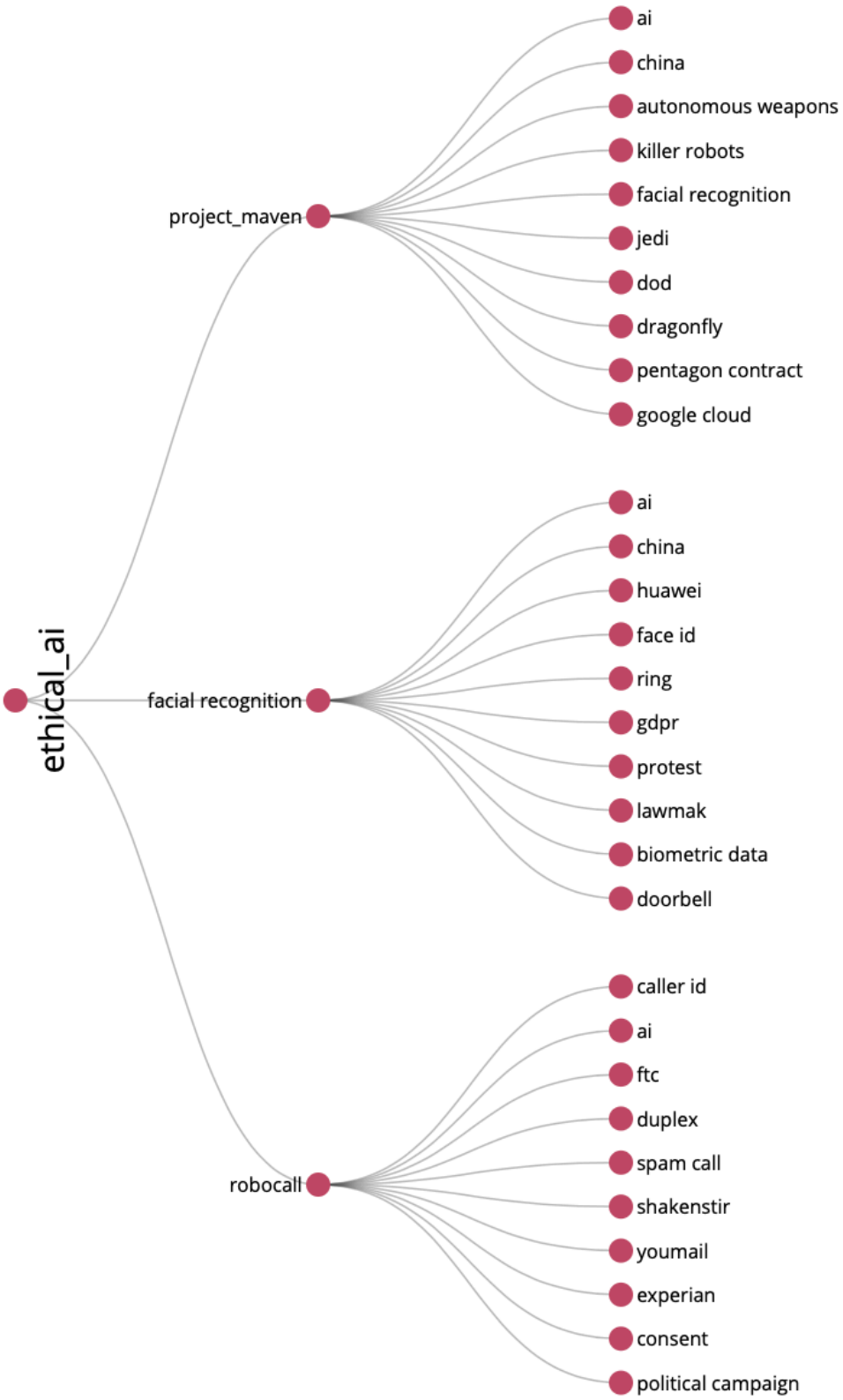


Table 22 Co-occurrences with most positive and negative sentiments: **facial recognition**

Most positive	Most negative
byton thinkpad voice assistant zoom ml	afr uighur metropolitan police cardiff hikvis

Table 23 Co-occurrences with most positive and negative sentiments: **project maven**

Most positive	Most negative
open ai democracy defence infrastructure DOD Jedi contract	lethal autonomous weapon system human oversight ban Amazon recognition

Table 24 Topic modelling: Ethical AI

Topic 1 Government (71.9 % of tokens)	Topic 2 Smartphones (18.1% of tokens)	Topic 3 Social media (5.9% of tokens)
ai google amazon data facebook government contract china facial recognition police	Apple phone app device facial recognition camera screen data display image	Facebook conservative google youtube Trump Twitter women republican political ad censorship
Topic 4 Robocalls (1.2% of tokens)	**Topic 5 Police (1.1% of tokens)**	**Topic 6 Airports (0.8% of tokens)**
Robocall carrier fcc at&t block spoof caller id scam ftc pai	smith ICE facial recognition Axon body camera law enforcement immigrant civil liberty CBP border	airport Brackeen law enforcement age verification DHS visa passport Amazon Go TSA boarding pass

6. Covid-19 analysis

In the second part of the analysis, we focus on the technological and social aspects of the current pandemic. Our aim is to support policy makers with providing up-to-date information on the various contexts of the crisis. We seek an answer to the questions:

- Which are the promising technological solutions (e.g. related to contact tracing)?
- Which are the most pressing technological challenges?
- What is the public perception of lockdown measures?
- Which are the trending research areas?

6.1 Topic identification

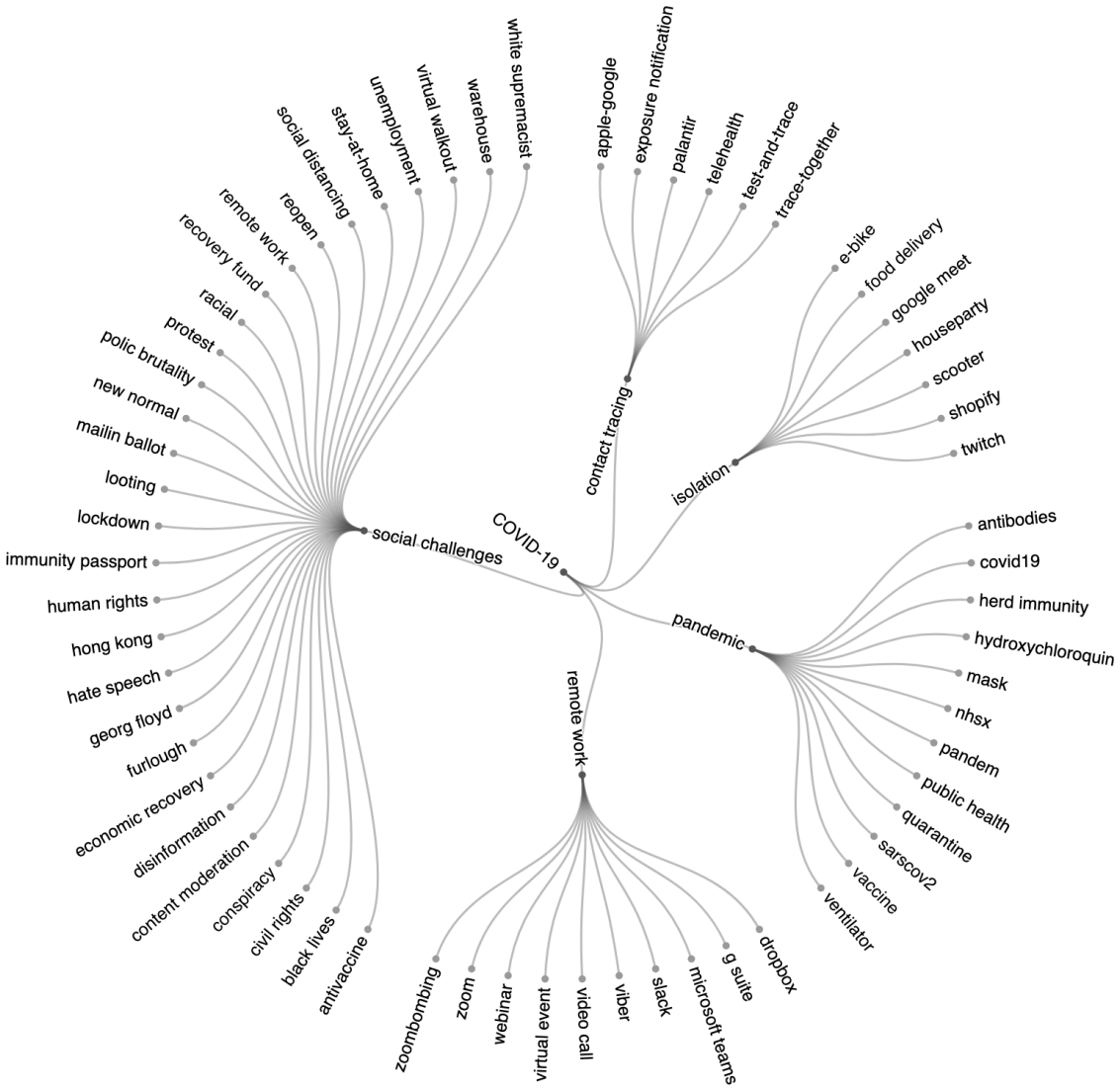
The trend analysis confirms that the majority of topics in news articles revolved around the COVID-19 pandemic. News reported on the medical aspects of the fight against the pandemic, as well as on the technological solutions that can help slow down the spread of the virus. The widely covered areas include contact tracing, apps for remote work, messaging and various services and products that support social distancing.

On the other hand, all major social challenges discussed in the previous sections remained relevant more than ever. With the worsening of the economy and growing social unrest, wide protests swept through the US. Social media is still riddled with misinformation and fake news, spreading conspiracy theories also on the novel coronavirus[36]. Employees of tech giants still protest against their employers, organising “virtual walkouts” during distance work[37].

[36] https://www.theguardian.com/world/2020/may/14/coronavirus-viral-video-plandemic-judy-mikovits-conspiracy-theories

[37] https://www.theguardian.com/technology/2020/jun/01/facebook-workers-rebel-mark-zuckerberg-donald-trump

Figure 15 Umbrella topics and identified keywords in news articles



6.2 Sentiment analysis

In the next step, we carried out sentiment analysis around five important COVID-19 related issues. For this purpose, we identified co-occurrences with the most positive and negative sentiment for each of the selected issues. The analysis was prepared using the VADER package[38].

First, we analysed sentiment around the PEPP-PT coalition - working on establishing a common standard for contact tracing solutions. This analysis was based on articles from the beginning of the pandemic. Next, we looked into a related issue, pertaining to joint efforts of

[38] VADER sentiment package is described in the first part of the report, see: 3.3.

Apple and Google to create a privacy-friendly contract tracing API. In the third analysis, we presented sentiments related to the Zoom app, which grew in popularity through the pandemic - but at the same time did not avoid controversies. In the fourth instalment, we examined the issue of disinformation, which is often considered a second pandemic[39]. In the last analysis, we presented sentiments around technological and political aspects of ventilators.

6.2.2 PEPP-PT

News articles in the wake of pandemic reported extensively on various tech projects related to tracking the movement of infected individuals. PEPP-PT is a coalition of researchers and tech experts across Europe developing a common standard for contact tracing solutions[40]. The sentiment in paragraphs covering PEPP-PT and other tech solutions has been positive or neutral.

A major angle of discussion is related to privacy and the fear of surveillance. Data protocols can be either "centralised", with data stored in a central location or "decentralised". Both approaches have advantages and disadvantages: in case of centralised systems more advanced data analysis is possible, but with a higher risk of data breach or violation of privacy[41]. Various national governments initially opted for centralised solutions, while the choice led to conflicts within the PEPP-PT project with researchers leaving, including Marcel Salathé[42]. Researchers preferring a decentralised data protocol published an own standard within the DP-3T project, an alternative for PEPP-PT[43]. Additionally, when Google and Apple joined forces to prepare an API for decentralised contact-tracing, national governments needed to reconsider their strategies[44].

The identified co-occurring terms include various actors of the recent news around PEPP-PT: researchers (Fraunhofer Heinrich Hertz Institute, École Polytechnique Fédérale de Lausanne, Michael Veale from Digital Rights & Regulation at UCL, Mayank Varia from Boston University[45]), companies (startup Arago with CEO Hans-Christian Boos[46], PocketCampus) the European Data Protection Supervisor Wojciech Wiewiórowski, government officials (Helge Braun), MEPs (Axel Voss[47]) and worldwide contact-tracing projects (the TraceTogether app in Singapore[48], Aarogya Setu in India[49]).

---

[39] https://www.nature.com/articles/d41586-020-01409-2
[40] https://www.pepp-pt.org/
[41] https://www.bbc.com/news/technology-52355028
[42] https://www.euractiv.com/section/digital/news/epp-cite-controversial-pepp-pt-as-example-for-single-european-covid-19-app/
[43] https://github.com/DP-3T/documents
[44] https://www.ft.com/content/7416269b-0477-4a29-815d-7e4ee8100c10
[45] https://www.zdnet.com/article/researchers-invent-method-to-track-coronavirus-through-smartphones-while-protecting-our-privacy
[46] https://hiroai.co/
[47] https://www.euractiv.com/section/digital/news/epp-cite-controverssial-pepp-pt-as-example-for-single-european-covid-19-app/
[48] https://www.euractiv.com/section/digital/news/epp-cite-controverssial-pepp-pt-as-example-for-single-european-covid-19-app/
[49] https://www.mygov.in/aarogya-setu-app/

Table 25 Co-occurrences with most positive and negative sentiments: **PEPP-PT**

Most positive	Neutral
Aarogya Setu	Hans Christian
proximity-tracking	Axel Voss
privacy first	Pocket Campus
Fraunhofer	decentralised model
Michael Veal	Schweisshelm
Wiewiórowski said	anonymised data
DP-3T	phone-tracking
Lausanne	Mayank
Helge Braun	gps location
central server	data-slurp

6.2.2 Apple-Google

The positive paragraphs report on the joint effort of companies to enable the use of Bluetooth technology and common API to help reduce the spread of the virus[50].

Neutral and slightly negative paragraphs focus on conspiracy theories related do COVID-19 tracing apps and privacy concerns related to certain apps regarding storing personal data on a central database[51] or sharing data with third-party services[52].

[50] https://www.apple.com/covid19/contacttracing/
[51] https://www.smh.com.au/politics/federal/privacy-advocates-raise-new-concerns-with-covidsafe-app-20200511-p54rwb.html
[52] https://gdpr.report/news/2020/06/08/privacy-north-dakota-contact-tracing-app-ends-data-share-with-third-party-services/

Table 26 Co-occurrences with most positive and negative sentiments: **Apple-Google**

Most positive	Neutral
google-apple api	conspiracy
decentralised protocol	privacy concern
privacy preserving	covidsafe app
trace app	test positive
french government	north dacota

6.2.3 Zoom

The analysed paragraphs fall into positive or neutral sentiment. The text snippets with higher score include competing solutions, the use-cases during the pandemic and useful privacy preserving features (e.g. background changing). On the other hand, the paragraphs with the lowest scores present the security and privacy issues that were unraveled recently, such as "zoombombing"[53]. With increased use of video chat solutions and higher scrutiny, competing app Houseparty also faced allegations related to security that were dismissed by the company (offering a bounty to identify the source of the "commercial smear"[54]).

Table 27 Co-occurrences with most positive and negative sentiments: **Zoom**

Most positive	Neutral
skype	zoombombing
slack	stamos
work remotely	end-to-end
background	trump
telehealth	zuckerberg

[53] https://www.theguardian.com/technology/2020/mar/27/trolls-zoom-privacy-settings-covid-19-lockdown
[54] https://www.zdnet.com/article/houseparty-app-offers-1m-reward-to-unmask-entity-behind-hacking-smear-campaign

6.2.4 Disinformation

The positive paragraphs include the efforts of social media platforms to tackle COVID-19 fake news[55]. Commission Vice-President Věra Jourová, who presented a communication on 'Tackling COVID-19 disinformation', is also mentioned.

The most negative news include conspiracy theories about vaccines[56], and the role far-right political movements and media play in spreading disinformation.

Table 28 Co-occurrences with most positive and negative sentiments: **Disinformation**

Most positive	Most negative
stamos	fox news
snap	vaccine
230	farright
jourova	tweet
content moderation	conspiracy theory

6.2.5 Ventilators

The positive fragments describe new mass-producible ventilators[57] and inventions allowing multiple patients to use a single ventilator[58].

The slightly negative and neutral paragraphs report on the struggles in the US to manage the pandemic.

[55] https://www.theverge.com/2020/3/16/21182726/coronavirus-covid-19-facebook-google-twitter-youtube-joint-effort-misinformation-fraud
[56] https://www.nature.com/articles/d41586-020-01452-z
[57] https://virginorbit.com/virgin-orbit-uci-and-ut-austin-design-new-mass-producible-ventilator-for-covid-19-patients/
[58] https://www.sciencedaily.com/releases/2020/06/200624100041.htm

Table 29 Co-occurrences with most positive and negative sentiments: Ventilators

Most positive	Most negative
virgin orbit	governor
face shield	fremont
nhs	fauci
petersen	chloroquine
recovery	trump

6.3 Topic modelling

Next, we analysed the topics of articles. We focused on articles that contain any of the terms in root form: "coronavirus", "covid-19" or "contact-tracing" (the list of terms is included in the Appendix).

Again, we use LDA to analyse characteristic words. So far, we did not consider time dynamics in topic modelling - we took all analysed articles and mixed them up in one great bag. On the other hand, the vocabulary of news articles is not static, but changes over time.
In order to examine trends, we used the dynamic topic modelling (DTM) approach, where the vocabulary of each topic evolves with time. The algorithm learns which are the most important topics, and how the characteristic terms changed in time. For this analysis, the algorithm explored topics for articles published every two weeks.

The table below presents the most characteristic words for the topics based on articles published in the last two weeks of June. The analysis shows that news articles extensively discussed the tech side of the pandemic: Topic 4 is related to contact tracing, while Topic 5 is reporting on the fake news crisis in the social media. Other major topics include Topic 6 on medical research, Topic 7 on the progress of the pandemic, Topic 8 on remote work,
and Topic 9 on the industry and economy.

Table 30 Topic modelling: Articles on Covid-19

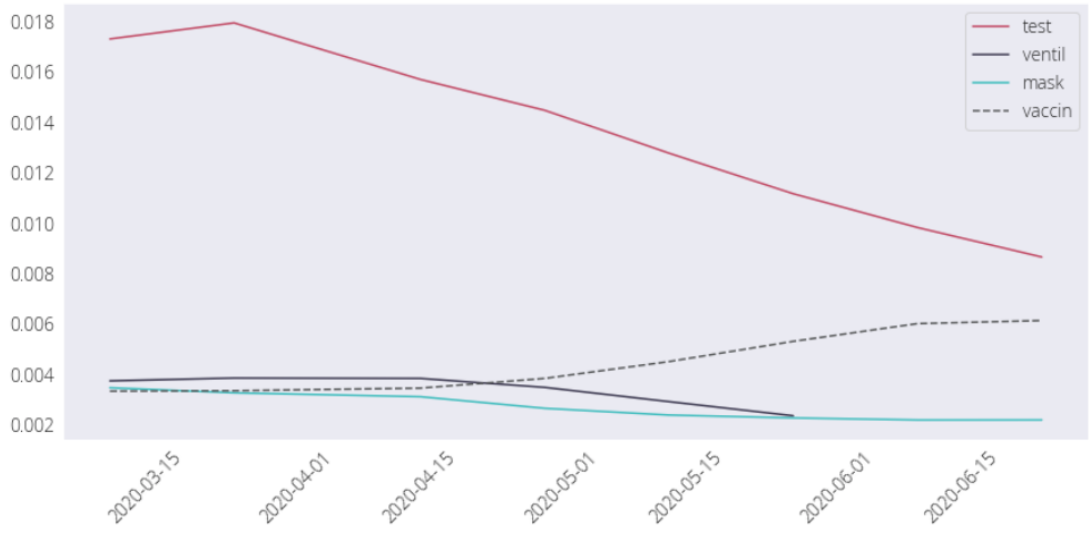
Topic 4 Contact tracing (11.4 % of tokens)	Topic 5 Social media (9.8% of tokens)	Topic 6 Research (9.5% of tokens)
app data government eu contact tracing technology	facebook twitter content google ad protest	research virus test infect patient vaccine

apple uk privacy google	trump remove misinformation tiktok	develop hospital model antibody
Topic 7 Pandemic (9.4% of tokens)	Topic 8 Remote work (8.4% of tokens)	Topic 9 Industry (7.9% of tokens)
case country trump pandemic mask death reopen new york spread china	app zoom google security microsoft attack release teams meet call	employee amazon worker tesla order delivery warehouse factory reopen staff

Besides analysing the characteristic terms in various periods of time, we can also evaluate changes in the frequencies of terms within the topics.

In the case of Topic 6 (medical research), we tracked the changes in term probabilities for the terms “test”, “ventilator”, “mask” and “vaccine” (in root forms, as in the figure). The analysis shows that frequency of ‘test’ gradually declined, while the role of vaccines increased. The importance of masks remained relatively stable, while ventilators disappeared from the most important 50 terms. These changes show us the focus of news articles: while during the spring discussions on ventilators and tests were more intensive, during the summer the focus shifted to vaccine research.

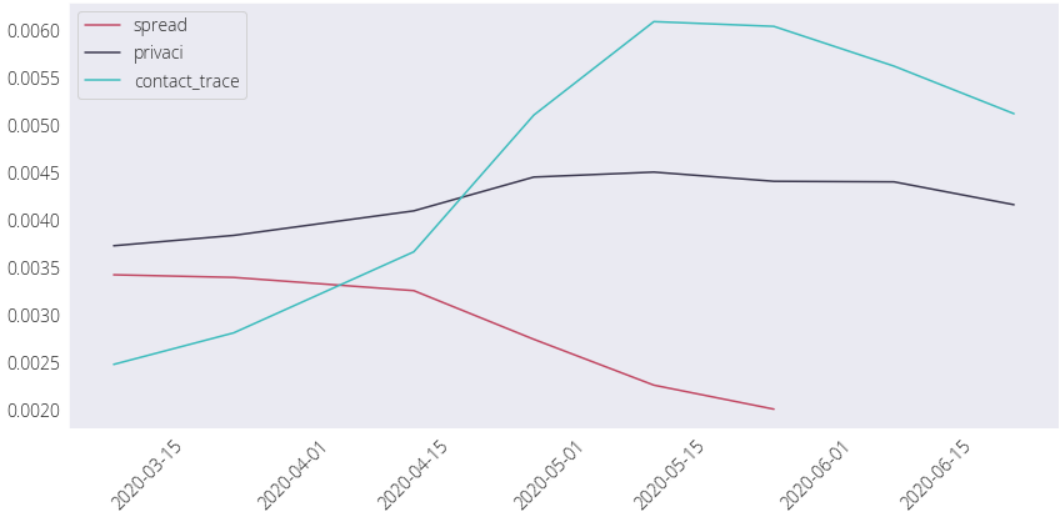
Figure 16 Dynamic Topic Modelling - medical research



In the case of Topic 4 (contact tracing), the changes show that more and more news reported on contact tracing until June, however, in the last week its role decreased. The focus on

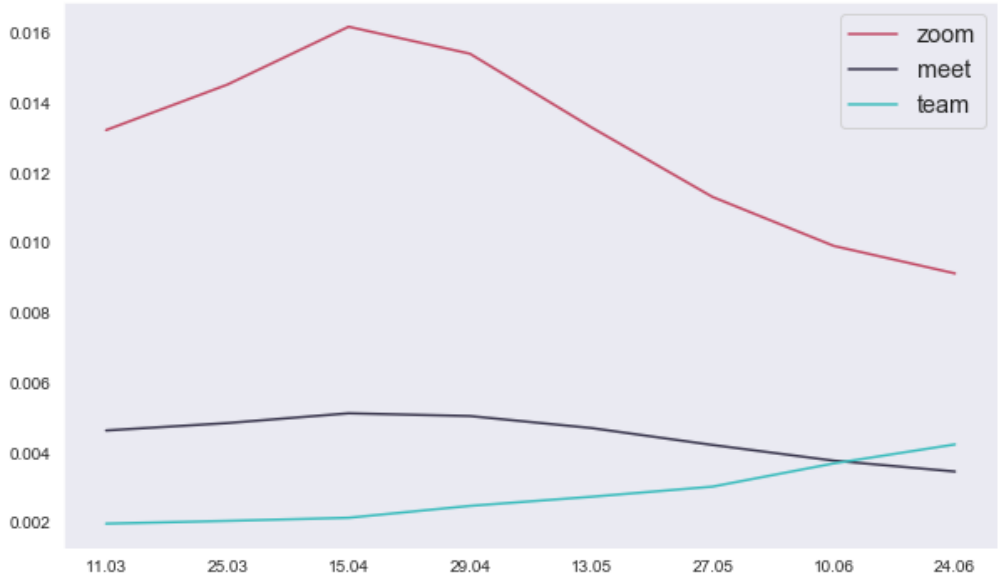
privacy has been relatively constant. Interestingly, the term “spread” disappeared from the most important terms in June.

Figure 17 Dynamic Topic Modelling - contact tracing



Finally, the changes in Topic 7 suggest that the wide interest around the Zoom platform declined. The term “team” potentially signals Microsoft Teams.

Figure 18 Dynamic Topic Modelling - apps



6.4 Mapping articles

We also mapped articles in a two-dimensional space. On the figure, each point represents an article, and articles covering the same subject are neighbours. In this analysis, we combined two methods: LDA and t-SNE. The colour of the point is assigned by LDA: articles with the same colour share a dominant topic.

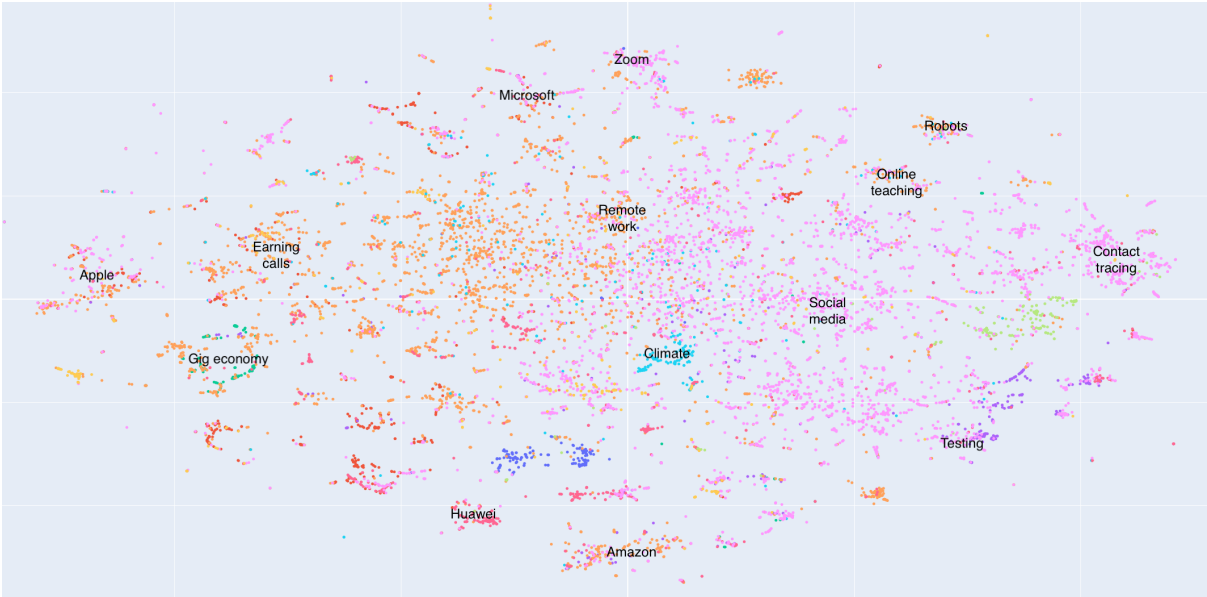
Usually, LDA creates a small number of very large topics that dominate most of the articles. In order to find greater variation and promote smaller topics, we assigned the dominant topic based on the average share of the topic[59]. While LDA is helpful in identifying wide topics, we are also interested in exploring more narrow areas. Therefore, we assigned articles in space using t-SNE.

On the map, articles covering the same subject are neighbours. We explored clear clusters on various topics that are marked on the figure. Users can also analyse the titles of articles in the interactive visualisations available on the website. Interesting areas include:

- Zoom: Security concerns (encryption issues, privacy and malicious users)
- Remote work: Company strategies, data security
- Microsoft: Surge in the use of Teams
- Robots: new ideas to implement robots during the pandemic (e.g. use in hospitals)
- Online teaching: tools to facilitate distant learning and concerns over the effectiveness of education
- Contact tracing: different apps and standards worldwide, privacy issues
- Social media: spread of Covid-19 misinformation, unrests in the US
- Testing: new diagnostic methods
- Climate: impact of the pandemic on social norms that affect climate change
- Amazon: protest of Amazon warehouse workers, also in Europe (France)
- Gig Economy: impact on workers, great importance of delivery services
- Earnings call: impact of the pandemic on tech giants
- Apple: closures and reopenings of stores, distant work of employees

[59] For each topic, we calculated the average share across all documents (global mean). The dominant topic of a document was determined by the deviation of the topic's share in the document from the global mean.

Figure 19 t-SNE map



6.5 GitHub

The open source community on GitHub has made attempts to explain the crisis and solve problems in a transparent way. GitHub users create repositories of code – projects containing versioned code. A majority of repositories are private, but those which are of most interest to us are public – they make GitHub the center of open-source collaboration. Anyone can contribute to the source code and other people can propose pull requests, which are accepted or rejected by maintainers. Users can also fork a repository – source code is copied from the original repository, giving the user full control over the new repository.

At the beginning of the pandemic, most repositories had been created in China. Early on, users based in Italy had a similar number of repositories created as larger EU countries, however, as the pandemic spread, users in the UK, Germany and France became more active.

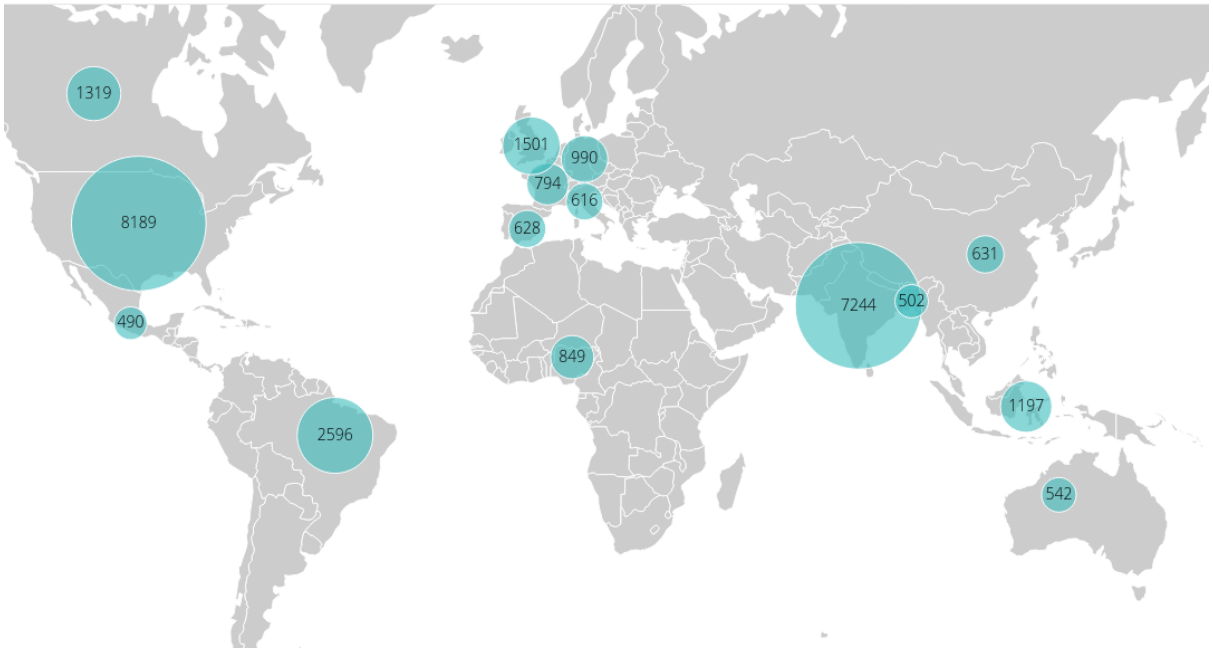
As of July, users in the United States and India have prepared the most repositories. American repositories are used and observed by the most users, and the most popular repository by far is by the Johns Hopkins University, providing data on the number of infections and deaths.

The low number of Chinese repositories does not mean a lack of interest in open source, but is rather caused by the disappearance of existing repositories, possibly forced by the state[60]. One of the most forked repositories, the 'wuhan2020-timeline,' was deleted in May[61].

[60] https://qz.com/1846277/china-arrests-users-behind-github-coronavirus-memories-page/
[61] https://github.com/Pratitya/wuhan2020-timeline/

Figure 20 Number of GitHub repositories with identified country as of 2020-02-10

Figure 21 Number of GitHub repositories with identified country as of 2020-07-13



In the wake of global pandemic, the open source community rose to the challenge. Hundreds of projects aiming mostly at mapping and predicting the spread of the virus have been created each day since the pandemic outbreak. However, after an initial exponential growth we have observed a slowdown in the number of created repositories in recent months. It is likely that most ideas have already been proposed and the community has moved to a more stable phase of working on existing projects.

In spite of slowing down activity in creating new repositories, the number of commits per week stays rather high, at about 100,000 per week. A commit is often not comparable to another – various projects may have different requirements. Still, the number of pull requests, used in projects with multiple branches, has decreased even more.

Figure 22 Daily number of COVID-19 related repositories created on GitHub

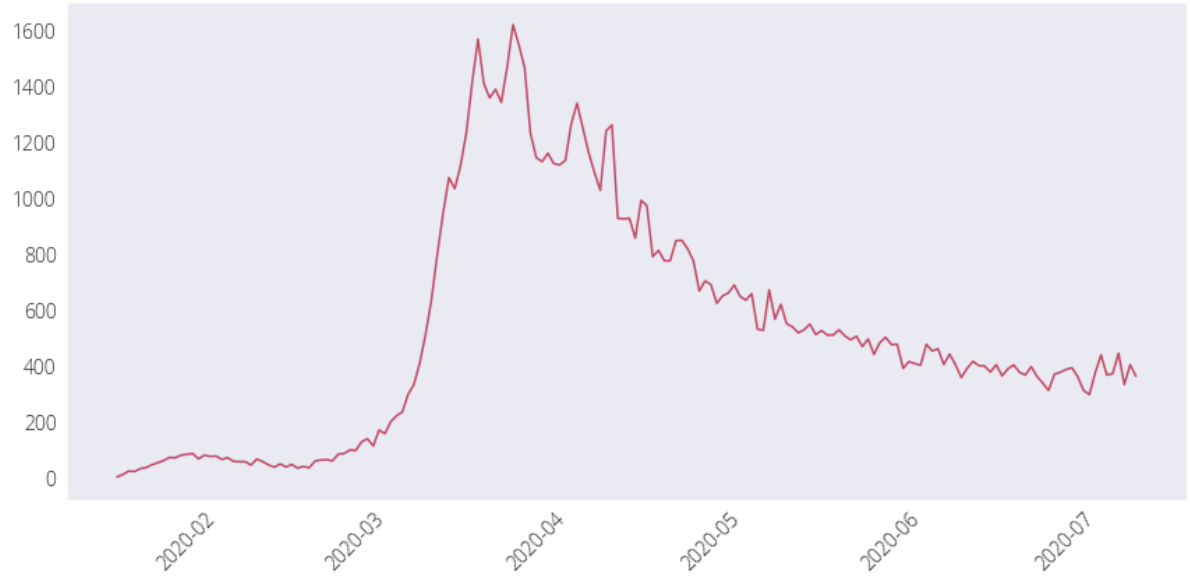


Figure 23 New commits added by week to GitHub repositories

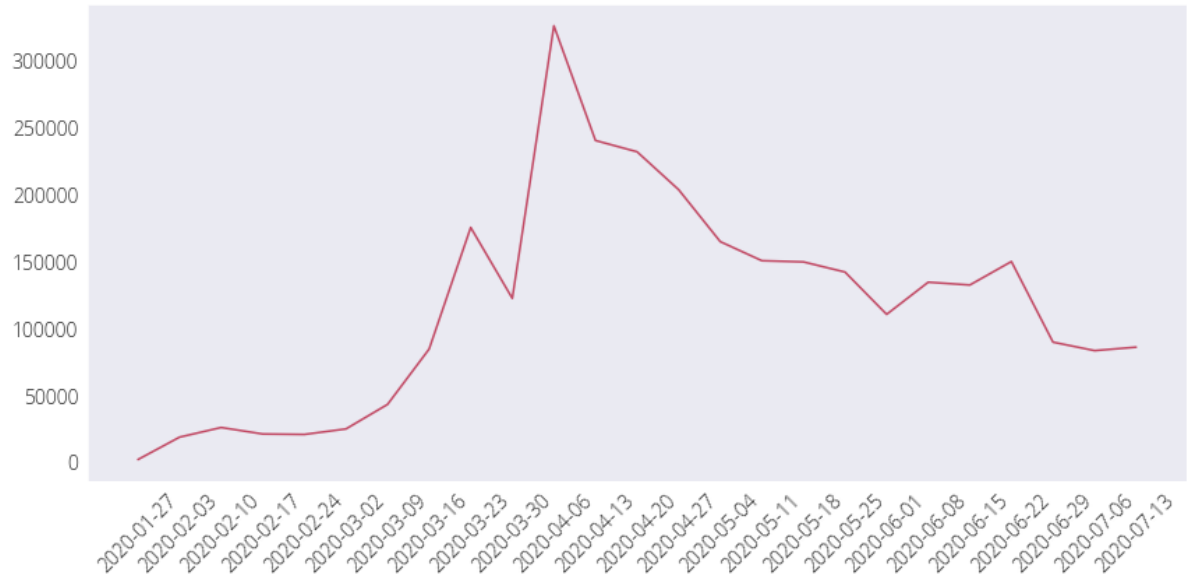
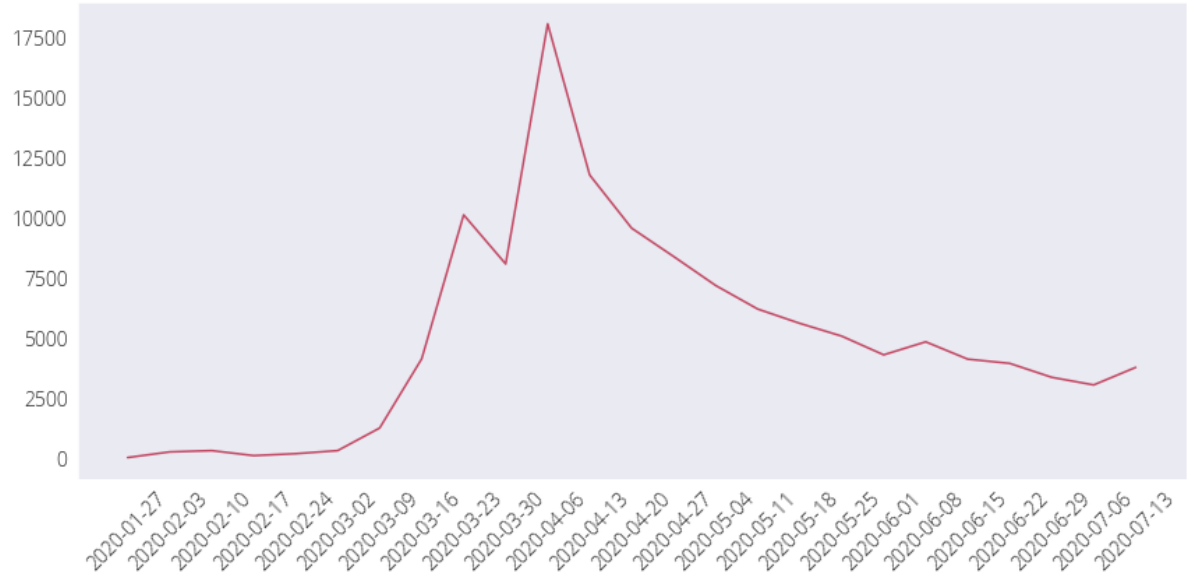


Figure 24 New pull requests by week in GitHub repositories



Most repositories are concerned with visualization: tracking the coronavirus and creating dashboards. Many projects analyse and model data using statistical tools, particularly for prediction. Is this work actually valuable? More analysis is required to assess the efficacy of such crowd-based knowledge.

Table 31 Most common words in GitHub repositories description, occurring at least 1000 times

word	count
data	10790
cases	4820
app	4818
using	4237
project	3913
analysis	2915
tracker	2684
api	2538
19	2445
application	2361
pandemic	2121
web	1973
simple	1913

world	1726
website	1715
dashboard	1699
information	1643
react	1631
python	1622
spread	1605
india	1603
model	1454
repository	1424
statistics	1414
based	1366
code	1364
tracking	1314
track	1295
people	1222
country	1194
time	1167
stats	1130
live	1128
number	1099
visualization	1096
help	1078
related	1046
learning	1031

We created an LDA analysis based on the project descriptions, an interactive version is available on the website[62]. The most common topic as found by LDA (see below) was tracking coronavirus, particularly in India (which explains a large number of repositories). The second one was charts, often using web technologies. Topic 7 are the repositories which may have the largest positive impact, particularly classification of vulnerable patients.

Table 32 Most salient terms for LDA topics in descriptions (lambda = 0 in the visualization on the website)

Topic 1 (14.2 % of tokens)	Topic 2 (11.5% of tokens)	Topic 3 (11.3% of tokens)
reactj	just	discord
visualis	attempt	telegram
globe	tableau	done
indian	evolut	retriev
recov	http	spain
página	javascript	graphql
modelo	awar	industri
dati	survey	morocco
pars	anoth	node.j
chartj	gama	canada
Topic 7 (8.9% of tokens)	**Topic 9 (8.5% of tokens)**	**Topic 10 (7.5% of tokens)**
angular	assign	aplicação
veri	demo	hook
classif	mask	ventil
react.j	face	webpag
mapa	blog	nigeria
chart.j	land	tuga
panda	scraper	curv
summari	caus	colombia
space	tabl	projet
aggreg	extens	good

6.6 Reddit

First, we examined the number of Reddit posts mentioning masks. The topic was sharply increasing in popularity until early April. The peaks can be attributed to the Czech response: both the social encouragement and government response. Among others, a wide campaign (#Mask4All) started by public figure Petr Ludwig in the social media mid-March became viral across Europe. Clearly, European countries can be world leaders in implementation, and well-applied principle of subsidiarity enables comparison of various approaches, while not discouraging cooperation.

[62] https://covid.delabapps.eu/model_github2007.html

Figure 25 Number of submissions about masks

The social environment puts a strain on mental health. Users of the r/anxiety subreddit in March posted almost twice as many comments on the subreddit daily, as in the pre-quarantine period. Interestingly, r/depression experienced a decrease in activity – the opposite of r/anxiety. Worryingly, r/SuicideWatch picked up activity after a fall in March. Theoretical knowledge of psychologists can be combined with real-time data to encourage politicians to make correct decisions and communicate them well: in a more calm or decisive way, depending on social needs.

Figure 26 Number of comments in r/Anxiety subreddit, 7-day moving average

Figure 27 Number of comments in r/SuicideWatch subreddit, 7-day moving average

Before the pandemic, unemployment was at a low level, with the American economy recording record-breaking job creation. Immediately, however, millions of people got laid off, which became visible in social media even earlier than in official unemployment data (in this chart we did not include terms such as "fired" or "laid off"). Apart from the social and economic challenge, it created various technological problems as well: mainframe-based systems (as used in several US states) were not prepared for such a wave of new registrations.

Figure 28 Number of comments containing the word “unemployment”, 7-day moving average

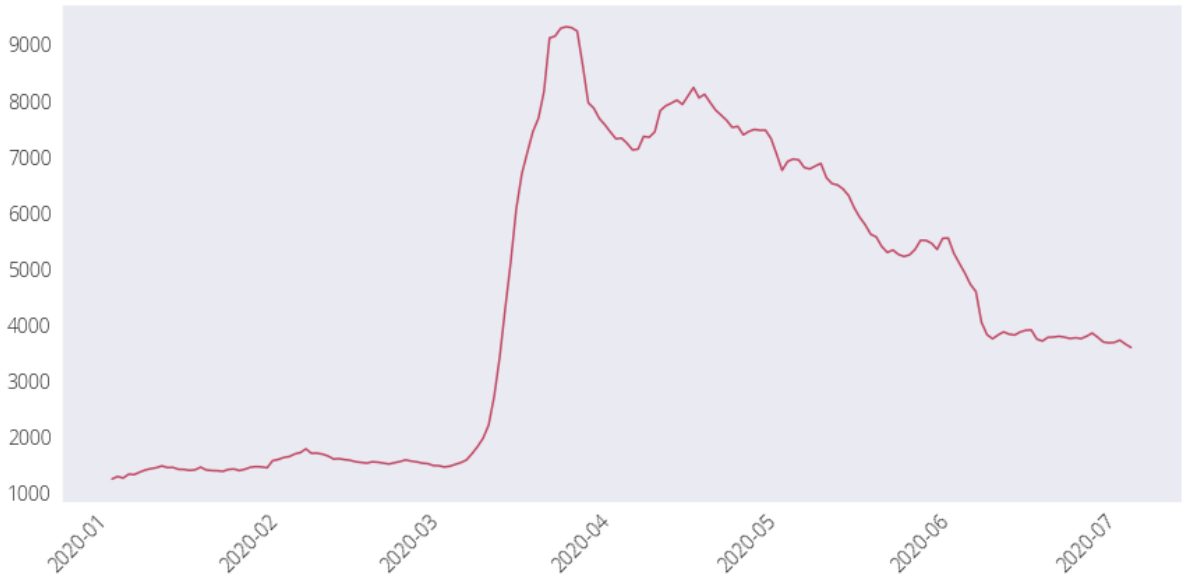


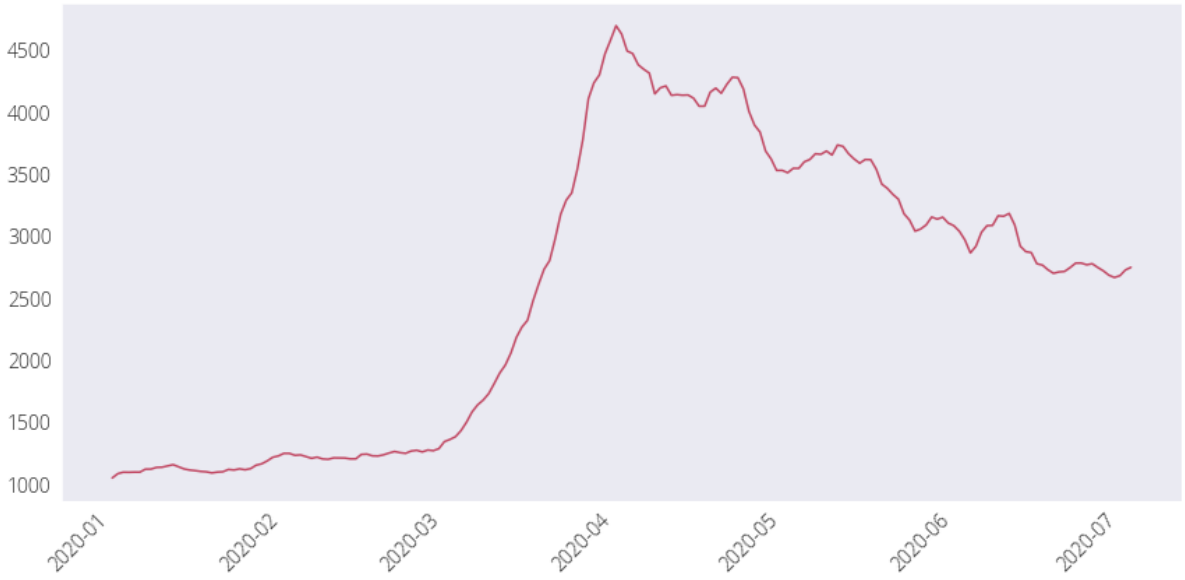
Table 33 Top 15 subreddits with comments containing the word “unemployment”

	subreddit	count
1	**politics**	45428
2	**wallstreetbets**	39832
3	**Unemployment**	33518
4	**Coronavirus**	29335
5	**AskReddit**	26576
6	**news**	18270
7	**personalfinance**	17140
8	**worldnews**	13485
9	**legaladvice**	9794
10	**Economics**	8309
11	**investing**	7409
12	**unpopularopinion**	7167
13	**AmItheAsshole**	6604
14	**smallbusiness**	6140
15	**pics**	5383

As an easy-to-use video conferencing app, Zoom has enjoyed increased popularity. It did not come without controversy, as Zoom routes some of its traffic by China[63] – a country notable for industrial spying[64]. Zoom has been called a "privacy disaster"[65], and has falsely claimed to use end-to-end encryption[66]. Nevertheless, its growth has been fairly stable and the number of comments regarding Zoom (regardless of context, in January and February its use might not have referred to the application) is still much higher than before the pandemic.

[63] https://techcrunch.com/2020/04/03/zoom-calls-routed-china/
[64] https://www.newsweek.com/china-involved-90-percent-economic-espionage-and-industrial-secrets-theft-1255908
[65] https://www.theguardian.com/technology/2020/apr/02/zoom-technology-security-coronavirus-video-conferencing
[66] https://techcrunch.com/2020/03/31/zoom-at-your-own-risk/

Figure 29 Number of comments containing the word “zoom”, 7-day moving average



Discussion on easing lockdown has been steadily increasing, although it is not as much of a topic as protests. It is connected to how people see lockdown: as an unpleasurable necessity, an infringement on their rights and an overreaction (like users of r/LockdownSkepticism), or a *too-little-too-late* measure?

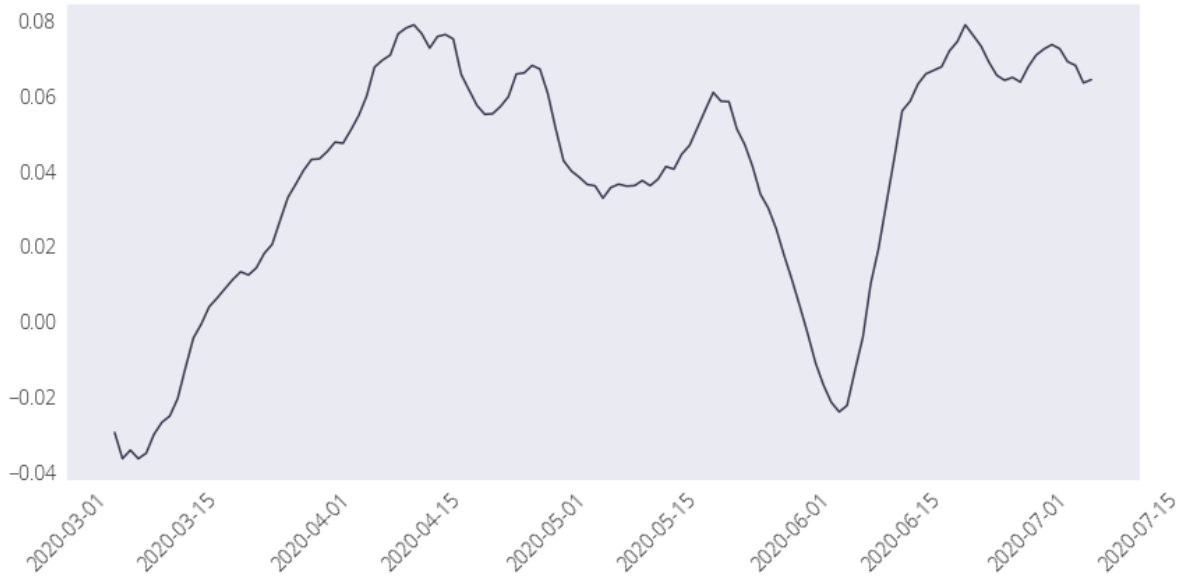
Table 34 Top 15 subreddits with comments containing the word “lockdown”,

	subreddit	count
1	**Coronavirus**	75920
2	**ukpolitics**	35486
3	**worldnews**	33237
4	**AskReddit**	32083
5	**DDnews**	20011
6	**LockdownSkepticism**	19485
7	**unitedkingdom**	18299
8	**india**	14543
9	**CoronavirusUK**	14402
10	**politics**	14136
11	**wallstreetbets**	13258

12	**newzealand**	11411
13	**conspiracy**	11119
14	**news**	9316
15	**China_Flu**	7754

The sentiment started from a level below 0: on average, comments were more negative than positive. As lockdowns were introduced in more and more places in March, the average sentiment value grew and stayed positive for the whole April. When countries began to ease restrictions, users worried it would result in a higher death toll, which makes the average sentiment lower once again. During Black Lives Matter protests – which have not been related to the lockdown, but took place during a period of restrictions – the sentiment value has initially dropped into negative values.

Figure 30 Average sentiment of comments with the word "lockdown", 7-day moving average



6.7 Working papers

Most common trending words in the arXiv dataset, despite not limiting the analysis to coronavirus-related articles, are generally related to the pandemic. Contact tracing is the most-studied topic. Other research has not ceased to be published: self-supervised learning, language models, and recommendation systems are trending as well.

In the dataset containing works from SSRN, a diverse set of problems can be identified, ranging from health-related (social distancing, incubation period, lockdown) to more common for social sciences: market, business or antitrust. Social sciences deliver a broader picture than news articles, which – even if focusing on social challenges – write rather about particular and immediate problems, such as small business closures or gig workers' plight.

Figure 31 Trending terms in ArXiv

Figure 32 Trending terms in SSRN

Conclusions

We have presented analyses for two periods of time, focusing on general trends in technology and society (January 2016 - December 2019) and the ongoing pandemic (January - June 2020).

In the first part, we have prepared deep dives for 8 key topics for the Next Generation Internet. The combination of text-mining methods provided us rich information on the most important news stories, concepts, actors, institutions and the relationships between them.

In the second part, we focused on examining the challenges posed by the COVID-19 pandemic. We explored in greater details such current issues as contact-tracing apps or remote work.

Besides news articles and working papers, we also analysed GitHub open source projects and Reddit discussions. The results showed that these data sources are helpful in tracking various social issues, including unemployment and mental health. Moreover, the analyses suggest Reddit can be also used to monitor public perception of lockdown measures.

While this report provides a description of results, the presented visualisations are static versions of the presentation available online at: https://fwd.delabapps.eu/ (first part) and https://covid19.delabapps.eu/ (second part). The results can be explored in interactive applications, enabling a deeper analysis.

Appendix

Table A1. Source weights

Source	**Weight**
Arstechnica	0.05
Euractiv	0.05
Fastcompany	0.05
Gizmodo	0.09
IEEE Spectrum	0.05
Politico Europe	0.05
Reuters	0.05
Techcrunch	0.09
Techforge	0.05
The Conversation	0.05
The Guardian	0.12
The Register	0.12
The Verge	0.09
ZDNet	0.09

Table A2. Terms used to filter articles in the deep dives

Topic	**Terms in root form**
Trustworthy information flows	{'rconspiraci', 'vaccin_misinform', 'moder', 'antivaccinerel', 'conspiracist', 'moder_team', 'detect_deepfak', 'conspiraci_theorist', 'conjur_misinform', 'counter_disinform', 'moder_by', 'youtub_moder', 'conspiraci_theori', 'moder_content', 'antivaccin_misinform', 'amplifi_disinform', 'harm_misinform', 'udeepfak', 'moder_resolut', 'polit_disinform', 'disinform_campaign', 'combat_disinform', 'misinform_problem', 'russian_misinform', 'rampant_misinform', 'moder_confid', 'antimisinform', 'antivaccin', 'spread_disinform', 'deepfak_detector', 'convinc_deepfak', 'spread_misinform', 'moder_practic', 'fight_misinform', 'onlin_disinform', 'antivaccin_group', 'former_moder', 'terrorismconspiraci', 'deepfak_pornographi', 'forum_moder', 'antivaccin_advoc', 'promin_antivaccin', 'combat_misinform', 'moder_polici', 'coordin_disinform', 'deepfak_detect', 'antivaccin_movement', 'misinform_spread', 'deepfak', 'misinform', 'content_moder', 'tackl_disinform', 'overmoder', 'conspiraci', 'antivaccin_activist', 'moder_process', 'contentmoder', 'fraud_conspiraci', 'moder_guidelin', 'russian_disinform', 'deepfak_video', 'disinform', 'curb_misinform', 'unmoder', 'moder_decis', 'misinform_campaign', 'disinform_tactic', 'peermoder', 'pornograph_deepfak', 'moder_tool', 'fight_disinform', 'human_moder'}
Blockchain and cryptocurrencies	{'blockchain_fintech', 'nonblockchain', 'stablecoin', 'decentr_cryptocurr', 'cryptocurrencymin_malwar', 'libra_associ', 'cryptoasset', 'blockchainasaservic', 'blockchainbas_platform', 'blockchain', 'blockchainrel', 'cryptocurr_balanc', 'cryptocurr_libra', 'mine_cryptocurr', 'blockchain_subsidiari', 'ibm_blockchain', 'blockchainbas_solut', 'blockchain_consortium', 'blockchainfocus', 'thirdlargest_cryptocurr', 'cryptocurr_mine', 'illicit_cryptocurr', 'blockchain_innov', 'blockchainpow', 'could_blockchain', 'cryptocurrencyblockchain', 'genevabas_libra', 'blockchain_expo', 'blockchain.com', 'corda_blockchain', 'libraton', 'cryptocurrencyfocus', 'blockchainequip', 'bitcoin_blockchain', 'asx_blockchainbas', 'cryptocurr_monero', 'coin_libra', 'cryptocurrencyback', 'cryptocurr_ether', 'blockchaindistribut', 'regul_cryptoasset', 'blockchainbas_payment', 'blockchain_technolog', 'cryptocurr_miner', 'cryptocurr_deriv', 'cryptocurr_wallet', 'public_blockchain', 'blockchainbas_chess', 'blockchainstyl', 'cryptocurr', 'socal_stablecoin', 'cryptocurr_trade', 'blockchainbas', 'cryptocurrencymin', 'cryptocurrencyexchang', 'facebookl_libra', 'utilis_blockchain', 'ethereum_blockchain', 'cryptocurrencybas', 'hackerturnedcryptocurr',

	'cryptocurr_theft', 'cryptocurr_trader', 'cryptocurr_ponzi', 'cryptocurrencyconnect', 'libra', 'libra_wallet', 'monero_cryptocurr', 'reuter_cryptocurr', 'blockchaindriven', 'blockchaintrad', 'libra_cryptocurr', 'stockcryptocurr', 'cryptocurr_scam', 'blockchainagnost', 'blockchainfuel', 'session_blockchain', 'blockchainbas_decentralis', 'libra_project', 'implement_blockchain', 'librastyl', 'blockchainasaservic_baa', 'cryptocurrency1', 'cryptocurr_exchang', 'facebook_libra', 'privat_blockchain', 'cryptocurrencyrel'}
Online privacy	{'privaci_violat', 'eprivaci', 'privaci_guard', 'oakland_privaci', 'privaci_polici', 'differenti_privaci', 'firm_nso', 'eus_eprivaci', 'apple.comprivaci', 'gdprlike', 'privaci_shield', 'pegasus_spywar', 'eprivaci_regul', 'privaci_scandal', 'facebookcambridg_analytica', 'privaci_guarante', 'eprivaci_propos', 'data_privaci', 'gdpr_compliant', 'austrian_privaci', 'privaci_dashboard', 'children_privaci', 'gdpr_legisl', 'onlin_privaci', 'privaci_control', 'gdpr', 'gdpr_mandat', 'nsos_pegasus', 'eus_gdpr', 'privaci_sin', 'privaci_rule', 'under_gdpr', 'gdprlimit', 'privaci_activist', 'privaci_badger', 'gdpr_general', 'privaci_regul', 'privaci_watchdog', 'set_privaci', 'privaci_concern', 'bloc_privaci', 'dataprivaci', 'privaci', 'privaci_practic', 'pegasus', 'gdpr_fine', 'gdprs', 'privaci_law', 'privaci_protect', 'proprivaci', 'privaci_enhanc', 'account_privaci', 'privaci_commission', 'broadbandprivaci', 'privaci_blunder', 'privaci_legisl', 'inform_privaci', 'stricter_privaci', 'strict_privaci', 'gdpr_regul', 'broadband_privaci', 'australian_privaci', 'privaci_profession', 'gdprstandard', 'privaci_advoc', 'privaci_implic', 'regul_gdpr', 'privaci_sandbox', 'privaci_right', 'privaci_nightmar', 'peopl_privaci', 'cambridg_analytica', 'privaci_hawk', 'euus_privaci', 'electron_privaci', 'analytica', 'privaci_safeguard', 'privaci_act', 'privaci_campaign', 'privaci_invas', 'privaci_intern', 'nso_group', 'privaci_conscious', 'pregdpr', 'privaci_set', 'gdpr_oblig', 'analyticalink', 'europ_gdpr', 'cambridg_analyticalink', 'analytica_improp', 'precambridg_analytica', 'privaci_laps', 'gdpr_complianc'}
Sustainable and climate friendly Internet	{'carbon_pollut', 'ewast_recycl', 'antipollut', 'lowemiss_vehicl', 'climat_catastroph', 'paleoclimatologist', 'microplast_pollut', 'climateneutr', 'greenhous', 'recycl_aluminum', 'recycl_mode', 'climat_activ', 'postclimateg', 'climat_central', 'climatedeni', 'climat_observatori', 'climateconcern', 'temper_climat', 'climatologist', 'climat_pledg', 'methan_emiss', 'co_2_emiss', 'co2_emiss', 'climat_action', 'pari_climat', 'climat_leadership', 'reduc_greenhous', 'ultralow_emiss', 'climat_scientist', 'harm_pollut', 'greenhous_effect', 'recycl', 'curbsid_recycl', 'emissionfre', 'pollutionspew', 'selfemiss', 'emissionscut', 'recycl_materi', 'aviat_emiss', 'thermal_emiss', 'greenhous_gass', 'climatemonitor', 'light_pollut',

	'netzero_greenhous', 'environment_pollut', 'address_climat', 'carbonemiss', 'renew_energi', 'climat_strike', 'climatedamag', 'gas_emiss', 'carbonpollut', 'climat_talk', 'climatevulner', 'zero_emiss', 'pollut_control', 'combat_climat', 'zeroemiss_vehicl', 'climat_chang', 'climat_striker', 'climatepollut', 'chang_climat', 'emiss_reduct', 'greenhous_gase', 'climaterel', 'biggest_pollut', 'recycl_facil', 'pollut', 'potent_greenhous', 'natur_climat', 'manmad_climat', 'climat_plan', 'climat_emerg', 'un_climat', 'reduc_emiss', 'cut_emiss', 'singlemiss', 'be_remiss', 'greenhous_emiss', 'climat_crisi', 'plastic_pollut', 'climatefuel', 'climat_denier', 'emiss', 'exhaust_emiss', 'emissioncut', 'lowemiss', 'emissionsfuel', 'emiss_test', 'cut_greenhous', 'rf_emiss', 'emiss_standard', 'warm_climat', 'climatethem', 'climat_apocalyps', 'climat_agreement', 'emissionspow', 'most_pollut', 'climat_activist', 'bemiss', 'ultralowemiss', 'climat_mobil', 'emiss_cheat', 'climat_shock', 'emissionsintens', 'climat_corp', 'pollut_fossil', 'emiss_target', 'emiss_line', 'climatefriend', 'industri_pollut', 'emiss_trade', 'worsen_climat', 'climat_changefuel', 'microclimat', 'carbon_emiss', 'air_pollut', 'artemiss', 'tougher_emiss', 'remiss', 'emiss_regul', 'recycl_ocean', 'global_climat', 'climaterel_disast', 'climat_justic', 'climat', 'climat_inact', 'batteri_recycl', 'netzero_emiss', 'pollut_account', 'climat_shift', 'breath_pollut', 'tackl_climat', 'runaway_climat', 'climateadjac', 'low_emiss', 'pollutioncurb', 'harm_emiss', 'recycl_program', 'climateg', 'climateinduc', 'pollut_industri', 'lesspollut', 'pollut_coal', 'emissionsfre', 'climat_impact', 'recycl_bin', 'climat_model', 'avert_climat', 'geopolit_climat', 'dioxid_emiss', 'climat_summit', 'climatolog', 'climatestrik', 'tailpip_emiss', 'deni_climat', 'climat_accord', 'zeroemiss', 'diesel_emiss', 'emiss_cut', 'youth_climat', 'climat_adapt', 'climat_scienc', 'greenhous_gas', 'recycl_plastic', 'climat_denial'}
Safer online environments	{'workplac_misconduct', 'spread_hate', 'antilgbtq', 'discriminatori_manner', 'bulli_hate', 'discriminatori_intent', 'hate_speech', 'hate_group', 'contain_hate', 'lgbtq', 'nondiscriminatori_basi', 'lgbtq_communiti', 'misconduct', 'nondiscriminatori_frand', 'hate_crime', 'misconduct_alleg', 'human_right', 'racial_discriminatori', 'whitesupremacist', 'hate_conduct', 'hatefil_communiti', 'hate_mail', 'discriminatori_practic', 'lgbtqi', 'hater', 'supremacistfriend', 'antisemit', 'hate_violenc', 'hatefil', 'alleg_discriminatori', 'hate_feminist', 'hate', 'nondiscriminatori', 'discriminatori_advertis', 'she_hate', 'remov_hate', 'discriminatori_hous', 'polic_hate', 'inher_discriminatori', 'white_supremacist', 'discriminatori_ad', 'misconduct_claim', 'antimuslim_hate', 'hate_content', 'everybodi_hate', 'hate_ideolog', 'hateabl', 'sexual_misconduct', 'organ_hate', 'antisemit_imageri', 'antisemit_comment', 'hate_messag', 'discriminatori',

	'supremacist', 'nondiscriminatori_term', 'most_hate', 'discriminatori_treatment', 'lgbtqaffliat', 'lgbtqia', 'lgbtq_advocaci'}
Democracy	{'democraci_technolog', 'elect_interfer', 'liber_democraci', 'matur_democraci', 'prodemocraci_protestor', 'repres_democraci', 'undermin_democraci', 'digit_democraci', 'antidemocraci', 'democraci', 'prodemocraci_protest', 'protect_democraci', 'sinorussianauthoritarian', 'polit_ad', 'countri_authoritarian', 'prodemocraci_activist', 'authoritarian_regim', 'western_democraci', 'polit_advertis', 'digit_authoritarian', 'our_democraci', 'defend_democraci', 'china_authoritarian', 'authoritarian_state', 'prodemocraci_movement', 'authoritarian_govern', 'authoritarian_rule', 'direct_democraci', 'authoritarian', 'digit_sovereignti', 'american_democraci', 'prodemocraci_demonstr', 'authoritarian_leader', 'polit_advert', 'prodemocraci'}
Market competition	{'antitrust_institut', 'huawei_equip', 'anticompetit_action', 'antitrust_rule', 'antitrust_divis', 'antitrust_violat', 'sweep_antitrust', 'anticompetit_merger', 'antitrust_scrutini', 'eus_antitrust', 'anticompetitv', 'anticompetit_law', 'antitrust_battl', 'antitrust_act', 'antitrust_ground', 'antitrust_approv', 'antitrust_law', 'antitrust_author', 'antitrust_probe', 'anticompetit_manner', 'antitrust', 'antitrust_suit', 'antitrust_concern', 'semiselfregul', 'potenti_anticompetit', 'antitrust_watchdog', 'anticompetit_provis', 'anticompetit_behavior', 'germani_antitrust', 'antitrust_fine', 'anticompetit_conduct', 'committe_antitrust', 'selfregul', 'antitrust_regul', 'selfregulatori', 'formal_antitrust', 'violat_antitrust', 'chines_telecom', 'europ_antitrust', 'hous_antitrust', 'anticompetit_practic', 'antitrust_enforc', 'antitrust_action', 'antitrust_lawsuit', 'tech_giant', 'alleg_anticompetit', 'chines_telecommun', 'industri_selfregulatori', 'anticompetit', 'anticompetit_behaviour', 'doj_antitrust', 'eu_antitrust', 'enforc_antitrust', 'antitrust_chief', 'antitrust_investig', 'antitrust_complaint'}
Ethical AI	{'bias_outcom', 'select_bias', 'bias_algorithm', 'age_verif', 'autonom_weapon', 'illeg_robocal', 'stop_robocal', 'robocal_block', 'spoof_robocal', 'against_robocal', 'polit_bias', 'antirobocal', 'gender_bias', 'unintend_bias', 'block_robocal', 'lessbias', 'independ_unbias', 'jedimobi', 'jedi_bid', 'fight_robocal', 'bias_toward', 'unbias', 'bias_against', 'human_bias', 'robocar', 'jedi_joint', 'unwant_robocal', 'autonom_weaponri', 'recenc_bias', 'institut_bias', '10000perrobocal', 'exist_bias', 'racial_bias', 'confirm_bias', 'robocallblock', 'implicit_bias', 'jedi_contract', 'facial_recognit', 'algorithm_bias', 'telephon_robocal', 'systemat_bias', 'societ_bias',

	'robocalyps', 'jedi_order', 'robocal', 'bias_discrimin', 'inher_bias', 'bias', 'unconsci_bias', 'jedi', 'system_bias', 'killer_robot', 'unmistak_bias', 'project_maven', 'biasmitig', '10bn_jedi', 'robocallcrush', 'jedi_cloud', 'unlaw_robocal', 'combat_robocal', 'jedi_evalu', 'robocorp'}
COVID-19	'covid19', 'covid19_pandem', 'covid', 'covid19_crisi', 'covidsaf', 'stopcovid', 'covid19_patient', 'covid19_outbreak', 'covid19rel', 'postcovid', 'postcovid19', 'covid19_vaccin', 'covidsaf_app', 'covid19_era', 'precovid', 'covid19_case', 'covid19_symptom', 'covid19_infect', 'precovid19', 'caus_covid19', 'covid19_contacttrac', 'covid_track', 'covid19_relief', 'covid19_epidem', 'covidrel', 'treat_covid19', 'covid_crisi', 'postcovid_world', 'stopcovid_app', 'my_covid', 'combat_covid19', 'australia_covidsaf', 'sever_covid19', 'contract_covid19', 'covid19posit', 'covid19_theme', 'covid19_impact', 'postcovid19_world', 'app_stopcovid', 'diagnos_covid19posit', 'covid19them', 'amid_covid19', 'postcovid19_workplac', 'covid19induc', 'covid19_respons', 'precovid19_level', 'covid19fuel', 'covid19_screen', 'covid19era', 'covidemda', 'covid19_communiti', 'noncovid', 'covidtailwind', 'covidavers', 'covid19deni', 'covid19tracer.ca', 'covidchalleng', 'covidinfect', 'covidrespons', 'covid19kil', 'covid19sputumsamplingswab', 'covidadapt', 'covidbot', 'covidpalooza', 'coronacovid', 'coronavirus', 'coronavirus_pandem', 'coronavirus_crisi', 'coronavirus_lockdown', 'novel_coronavirus', 'coronavirusrel', 'coronavirus_contacttrac', 'see_coronavirus', 'postcoronavirus', 'coronavirus_infect', 'coronavirus_vaccin', 'coronavirus_relief', 'coronavirusinduc', 'precoronavirus', 'coronavirus_task', 'coronavirus_contactstrac', 'coronavirus_sarscov2', 'coronaviruss', 'postcoronavirus_world', 'coronavirus_aid', 'coronavirus_busi', 'coronavirusl', 'coronavirus_misinform', 'coronavirusdriven', 'coronavirus_case', 'coronavirus_resourc', 'other_coronavirus', 'bat_coronavirus', 'ongo_coronavirus', 'coronavirustrac', 'coronavirusrel_restrict', 'coronavirusfuel', 'coronavirus_test', 'techrepubl_coronavirus', 'coronavirus_is', 'coronavirusrel_lockdown', 'cnet_coronavirus', 'coronavirusbatt', 'anticoronavirus', 'coronavirusthem', 'coronavirus_cure', 'coronavirus_respons', 'coronavirusrel_scam', 'coronaviruscur', 'coronaviruspandem', 'coronavirusprompt', 'dead_coronavirus', 'coronavirus_origin', 'coronavirus_fear', 'coronavirus_outbreak', 'contact_trace', 'contact_tracer'