D1.2: Visualisations of key emerging technologies and social issues

Work package	WP1: Topic identification
Task	1.2: Continuous identification of new technologies and trends 1.3: Topic co-occurrence analysis: where do technology and social issues meet? 1.4: Network Analysis through Topic modeling & deep learning algorithms
Due date	30.06.2019
Submission date	
Deliverable lead	DELab
Dissemination level	Public
Nature	Report
Authors	Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, Maciej Wilamowski
Version	
Reviewers	
Status	

Acknowledgement: This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme under grant agreement Nº825652**

NGI FORWARD

EXPLORATIONS IN NEXT GENERATION INTERNET

View the interactive presentation to see the visualisations and overview of our main findings.

1. Introduction

Technological development is strongly shaping our societies, affecting virtually all aspects of life. With the growing role of technology, even minor imperfections can create huge challenges: from algorithmic biases to the collection of personal data, negative externalities can easily reach a critical scale. In such a quickly changing environment, there is a great need to support informed policy-making, decreasing the lag between technological changes and regulatory responses.

Besides mitigating risks, it is also crucial to facilitate the development of Europe's human-centric digital economy. The identification of the most promising emerging technologies provides a guidance for further research and investments in the EU tech industry.

The major aim of this empirical analysis is to provide tools to map and explore the most important social challenges and emerging technologies. The report presents a methodology that can be implemented for various topics and sources, enabling a data-driven agenda-setting for policy, with problem recognition, definition and selection.

The text mining analysis is based on a novel dataset of technology news articles and academic working papers. The two types of sources are complementary to each other, representing different layers of information: while news articles are covering current affairs, academic research provides detailed insights into narrow fields. Therefore, news articles are suitable to identify general trends, while working papers help to pinpoint their most important aspects and areas. Working papers were chosen over published academic articles due the long publishing times of peer-reviewed journals: while working papers are mostly works in progress, presenting the newest results, published papers are often presenting research from previous years.

Using web-scraping tools, more than 213 thousand articles, and more than 139 thousand working papers were collected. The sources include 14 major English-language technology websites from the US, EU and Australia, and 2 working paper repositories, covering both social and STEM (science, technology, engineering, and mathematics) sciences.

First, we identify emerging topics based on the analysis of term frequencies. The demonstrated method enables to highlight the most trending technology-related topics,

without any prior assumptions or shortlists. The identified terms serve as input for further analysis. The connections between terms, e.g. between emerging social issues and technologies are explored using co-occurrence analysis. In order to track the public perception of issues and identify the positive and negative sides of selected topics, sentiment analysis was performed. To additionally verify our results, topic modeling was prepared using Latent Dirichlet Allocation, a complementary approach to our topic identification strategy.

Therefore, the combination of these 4 techniques provides a guide, mapping trending topics, establishing the relationships between them, identifying key actors and institutions, and also tracking changing public perception.

To demonstrate the potential of the methodology, deep dives are prepared for 8 umbrella topics, chosen from the most trending technologies and issues:

- Artificial intelligence and machine learning
- Internet of Things
- Blockchain and cryptocurrencies
- Quantum computing
- Internet regulation
- Social media and content crisis
- Market competition
- Chinese tech sector

Following the introduction of the sources used, the methodology is explained in section 3. Section 4 presents the identified emerging topics. The remaining sections focus on the deep dives, and the report ends with conclusions.

The presented results are available in the form of interactive visualisations at https://fwd.delabapps.eu/

2. Sources

This forecasting exercise is based on the analysis of information stemming from two important stakeholder communities: journalists covering technology news, and academia. 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. The methodology combines these two perspectives, providing a more holistic insight into the development of Internet technologies.

Figure 1. presents the 14 selected online media sources, including the number of collected articles and country of origin. These sources were chosen, because:

- they represent various geographical areas
- provided early signals on currently available technologies (e.g. VR, IoT) in the past
- the articles are publicly available.

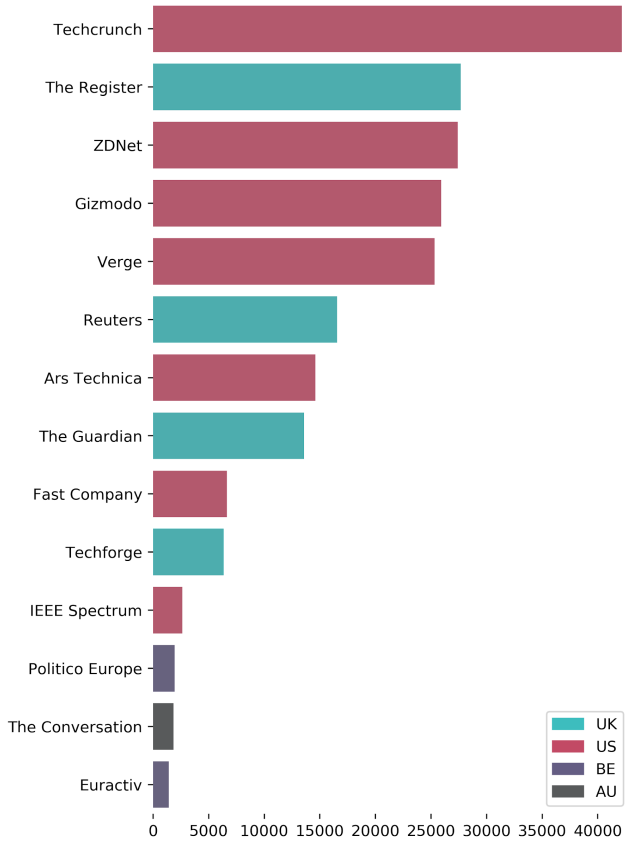
In the case of working papers, works from two repositories are collected: ArXiv (STEM sciences) and SSRN (social sciences).

ArXiv is owned and managed 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: AI, Computation and Language, Cryptography and Security, Data Structures and Algorithms, Human-Computer Interaction, Information Retrieval, Networking and Internet Architecture, to name only a few.

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 data sets of the sources are: online tech media (213 000 articles), working papers on computer science (118 000) and working papers on social sciences (20 000). The data has been collected for a period of 40 months (between 2016-01-01 and 2019-04-30). Besides article and working paper metadata (name of the author, publication date etc.), the plain text of the media articles have been collected, and the abstracts of the working papers. The data collection process and data structure have been described in detail in the deliverable D1.1 (Documentation of the database). The data sets prepared during the project are available at a Zenodo repository (https://zenodo.org/communities/ngi_forward/).

Figure 1. Number of media articles per source and country of origin



3. Methodology

3.1 Topic identification

We begin the analysis with the identification of trending topics in online news and working papers. Based on observing changes in term frequencies over time, the most trending terms are revealed. The methodology does not require any assumptions or the preparation of suspect short-lists, facilitating the discovery of unexpected but highly relevant areas.

For each month, the number of occurrences of each term was counted in each article separately. Then, these term counts for all articles from a given source and published in a given month was added up and the sum was divided by the number of such articles. The results (mean term occurrence per article in a given source and month) were averaged with predefined weights, resulting in a single number for each term/month pair. The weights are provided in the Appendix.

Then, a best-fitting regression line has been identified using ordinary least squares method, where the x axis contained numbers from 1 (first month) to 40 (last month), and the y axis – respective values for the term/month pair. The regression coefficient has then been divided by the mean frequency of the term, providing a normalised coefficient. The normalised coefficient is 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 0.025, 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.

Following the removal of stopwords and words not growing fast enough, the article terms were sorted by the regression coefficient. The top 20 most trending unigrams and bigrams are included in the Appendix. The top 1000 trending words and bigrams were reviewed, and the relevant terms for further analysis were selected.

Following the selection of relevant trending terms (172 unigrams and bigrams), they were organised into wide umbrella topics. These topics are presented in Figure 2 (interactive visualisation available online).

3.2 Co-occurrence analysis

The analysis of term frequencies served as an automated method to filter out the most relevant terms from the text corpus, providing an overview on the most important topics in the tech world. The next step is to explore selected topics deeper and establish the relationships between trending terms. The analysis of co-occurrences enables us to find which emerging terms were most often mentioned together in the same article, hence finding most relevant pairs of expressions. In the case of tech news, such method can be used to identify the areas where a technology is applied, or connections to regulatory issues.

First, the terms of interest are chosen among the identified trending words (e.g. machine learning). Next, co-occurring terms have been selected out of the top 15000 most trending terms. For each trending term, the number of its occurrences in articles containing the term of interest (like machine learning) has been counted for each source. An average index is calculated from the sources, using the same weights as in regression part, further normalization adjusts the values relative to the number of the trending term's occurrences (the values are divided by the square root of the trending term's frequency).

3.3 Sentiment analysis

Media articles are often polarising, and public perception of particular technologies and related regulatory and social issues may evolve over time. Therefore, changing sentiments of trending topics are examined. Additionally, news story involve positive and negative actors and relations. Analysing the sentiment of articles with co-occurring words, the different sides of debate can be identified.

Sentiment is calculated using the VADER package (Valence Aware Dictionary and SEntiment Reasoner)[1]. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in short texts.

For the main analysis of terms, all paragraphs in articles containing the given term are modified to exclude this term and assigned a score between -1 (most extreme negative) and 1 (most extreme positive) by VADER. Removal of terms is meant to exclude sentiment of the term itself, because the term may not be emotionally neutral, e.g. when some technologies or companies attempt to solve a negative issue. In such case, the neighbourhood's scores would be positive, but the negative term would bring the paragraph's score down. Two analyses will be presented: the average sentiment for each month, and co-occurrences with the most positive and negative sentiment. In the case of the latter, for each term, the 100 most co-occurring terms have been selected and sentiment computed for each paragraph modified once again by removing both the analysed and co-occurring terms. Terms present in most negative and most positive (on average) paragraphs are then extracted.

3.4 Topic modeling

The final element of our analysis is topic modeling with the use of LDA algorithm. The aim of this task is to automatically discover latent themes present in tech media, social science and STEM working papers.

Topic modeling refers to a combination of algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents (...) topic models can organize the collection according to the discovered themes"[2]. The most popular method of topic assignment is the Latent Dirichlet Allocation (LDA) model[3], 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. Being an unsupervised machine learning algorithm, LDA does not require providing a training dataset nor manual coding of the document. 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[4]. Instead of beginning the analysis with a predefined set of terms and codes derived from domain expertise, the researcher specifies the number of topics that the algorithm is supposed to find.

For each topic, the percentage of tokens (strings of characters) is presented. The higher percentage of tokens indicates a higher prevalence of a given topic.

---

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[2] D. M. Blei, Probabilistic topic models (article), Communications of the ACM (2012).

[3] 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).

[4] D. M. Blei, Topic Modeling and Digital Humanities, Journal of Digital Humanities (2012).

4. Topic identification

4.1 Online news

We begin the analysis with identification of trending terms in news articles. Based on the changes in term frequencies, the most trending terms relevant to NGI are selected. Figures 3. and 4. present the identified terms that reveal the most important technologies and social issues.
We grouped the identified terms into the following wide umbrella topics: AI, IoT, cloud and edge computing, robots, 5G, autonomous vehicles, blockchain and cryptocurrencies, quantum computing and mobile devices. The results show that not only widely used terms were captured (e.g. 5G infrastructure, AI startup, cloud computing, IoT technology), but also domain specific ones (e.g. 5g: mmwave, AI: pytorch, tensorflow, GPT-2, cloud computing: CNCF, IoT: co-location IoT). Another desired feature of the methodology is the lack of buzzwords of the past, such as the term “Big Data”.

Besides technologies, the following trending social issues were identified: privacy, political influence, #metoo, internet regulation, cybersecurity, content crisis, competition policy, Chinese tech sector, technology workers. Many of the umbrella topics are strongly related to each other, as in the case of content crisis and political influence, or online privacy and internet regulations.

Figure 3. Umbrella topics and identified keywords: Emerging technologies

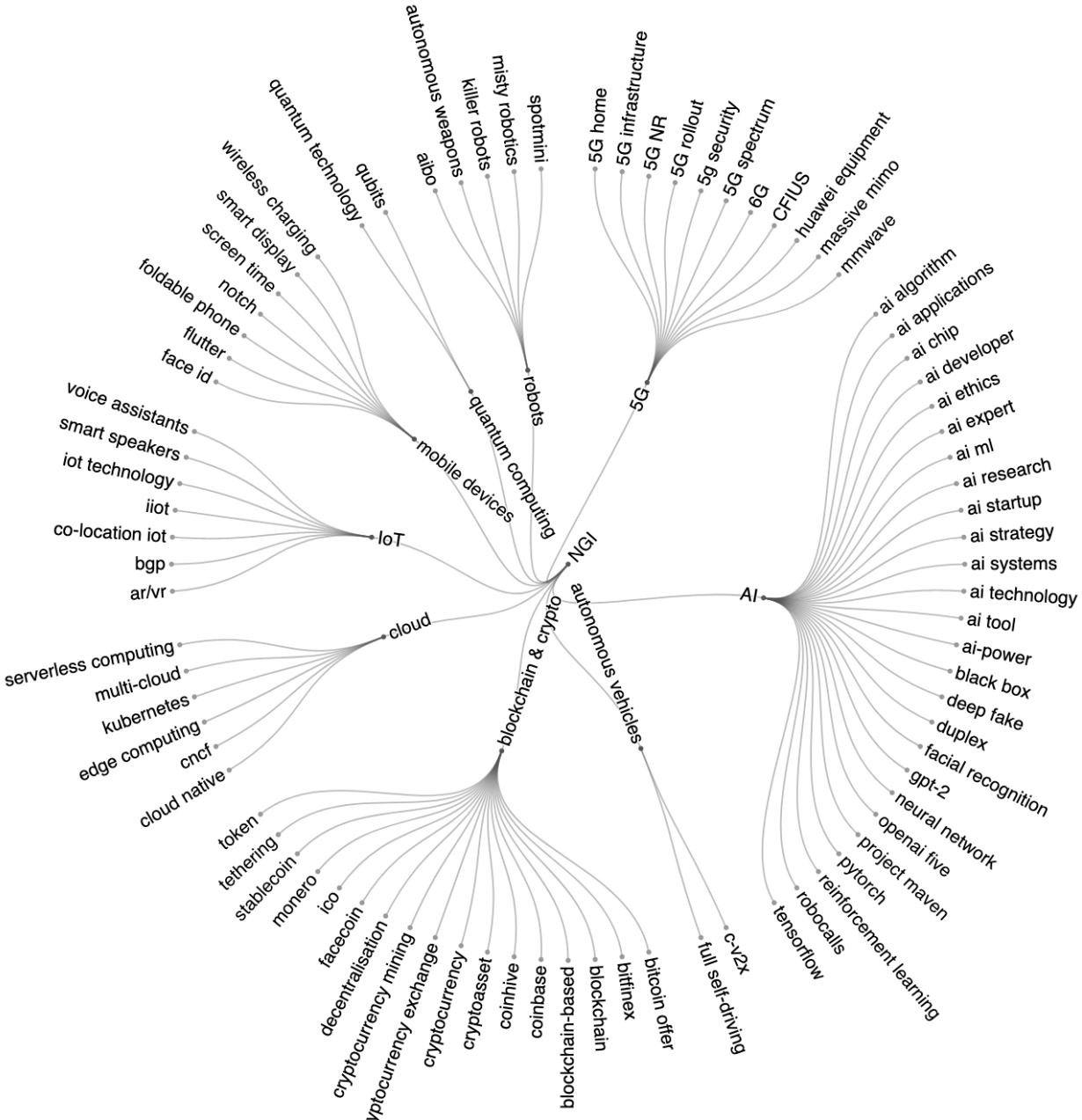
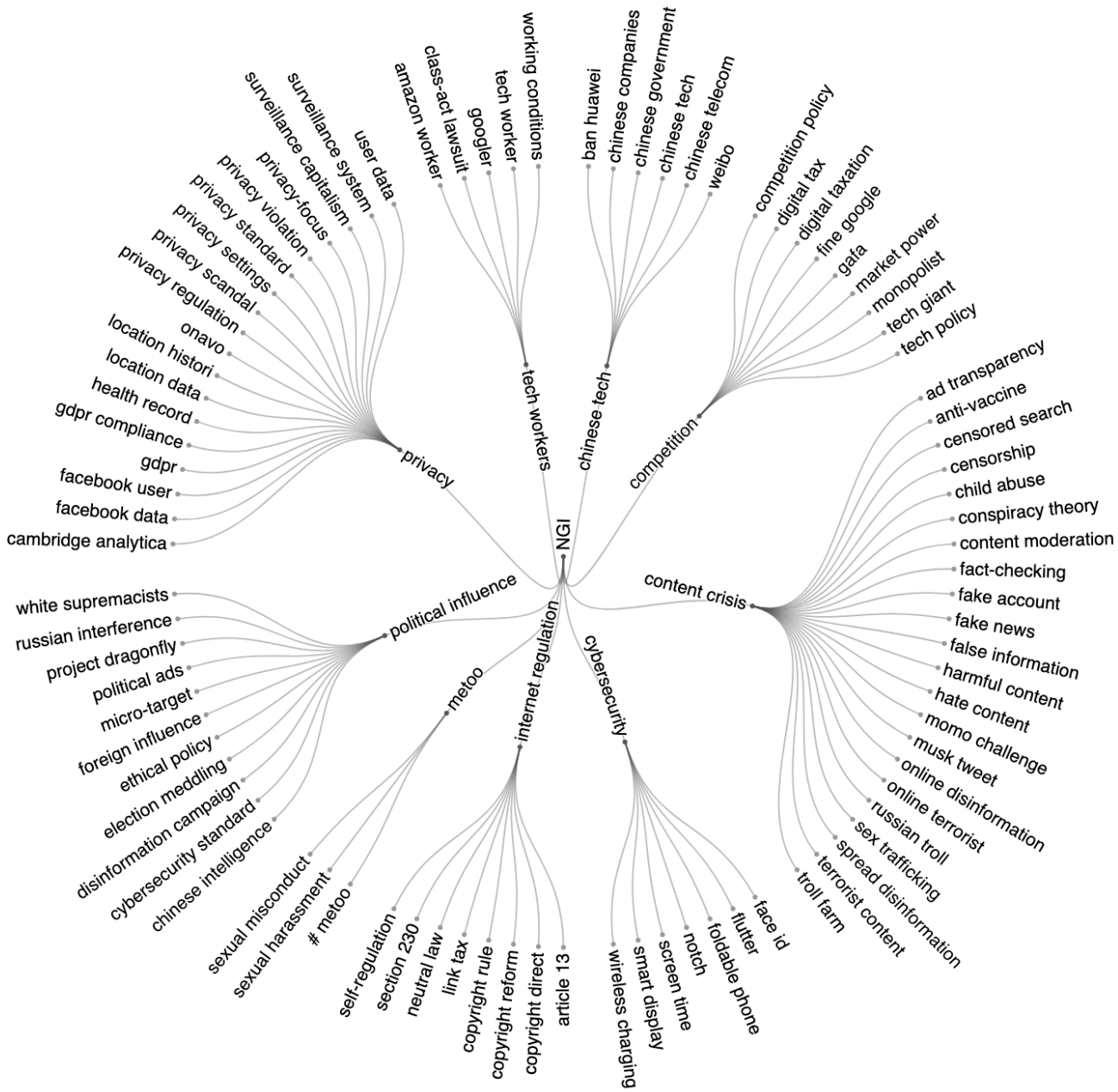


Figure 4. Umbrella topics and identified keywords: Relevant social challenges



4.2 Working papers

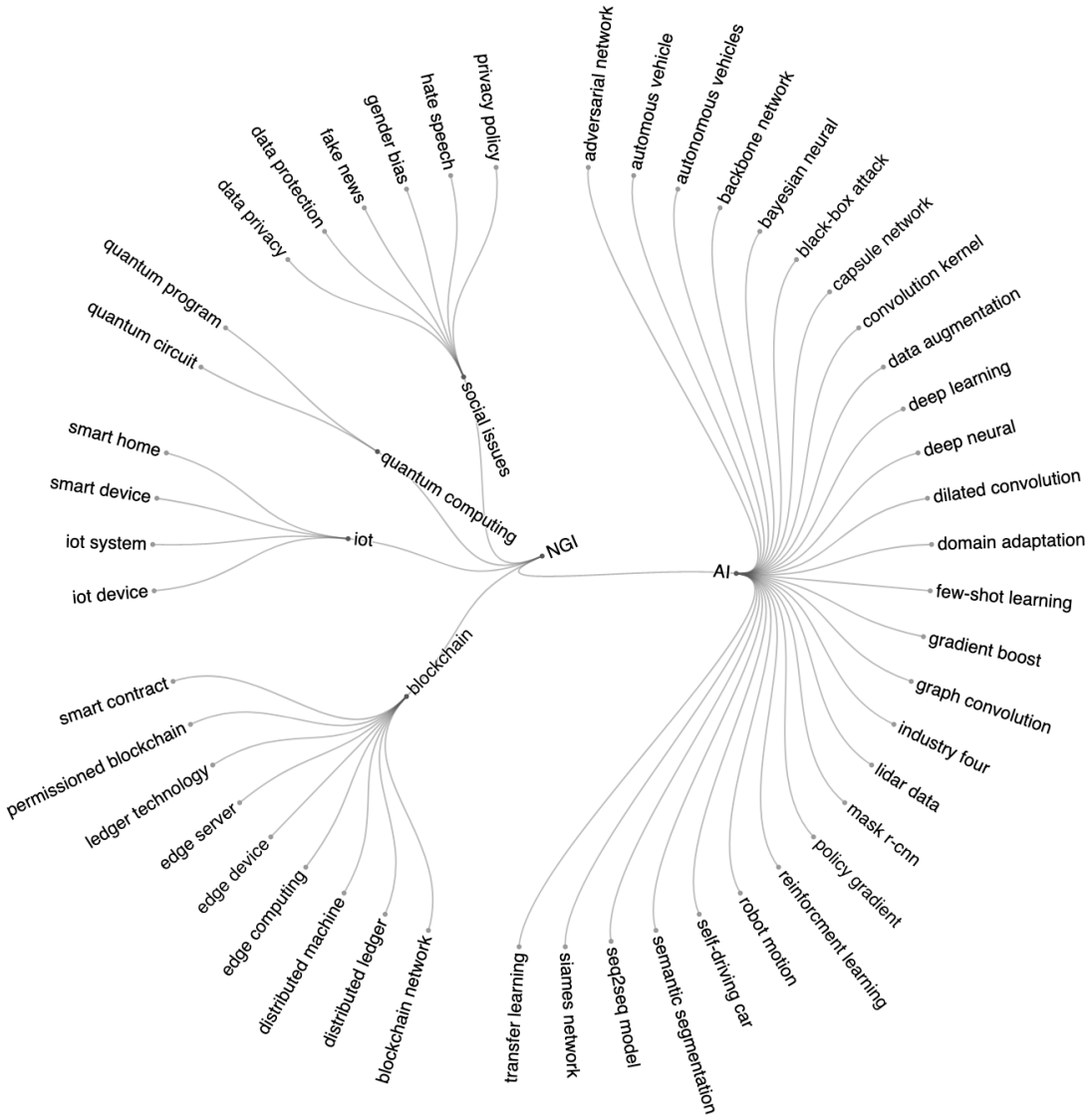
Next, the abstracts of papers in Arxiv and SSRN are analysed with the same method. Similarly to news articles, relevant terms are selected among the most trending unigrams and bigrams.

In the case of Arxiv, various AI and ML methods are strongly trending, including methods based on neural networks, reinforcement learning and gradient boosting. The analysis of trends enables the identification of novel areas within deep learning methods, such as generative adversarial learning or convolutional networks. Moreover, the method also reveals more narrow fields within these research areas, such as few-shot learning methods.

Besides the algorithms, the analysis also shows examples of practical implementation, such as autonomous vehicles, self-driving cars, robot motion or Industry 4.0.

Research interest has also been growing in the field of decentralised technologies, such as blockchain, distributed ledgers or edge computing. Other trending areas include quantum computing and internet of things devices. Finally, while computer science and natural science researchers form the main community of Arxiv, several social challenges were identified in the papers, including fake news, online privacy or gender bias.

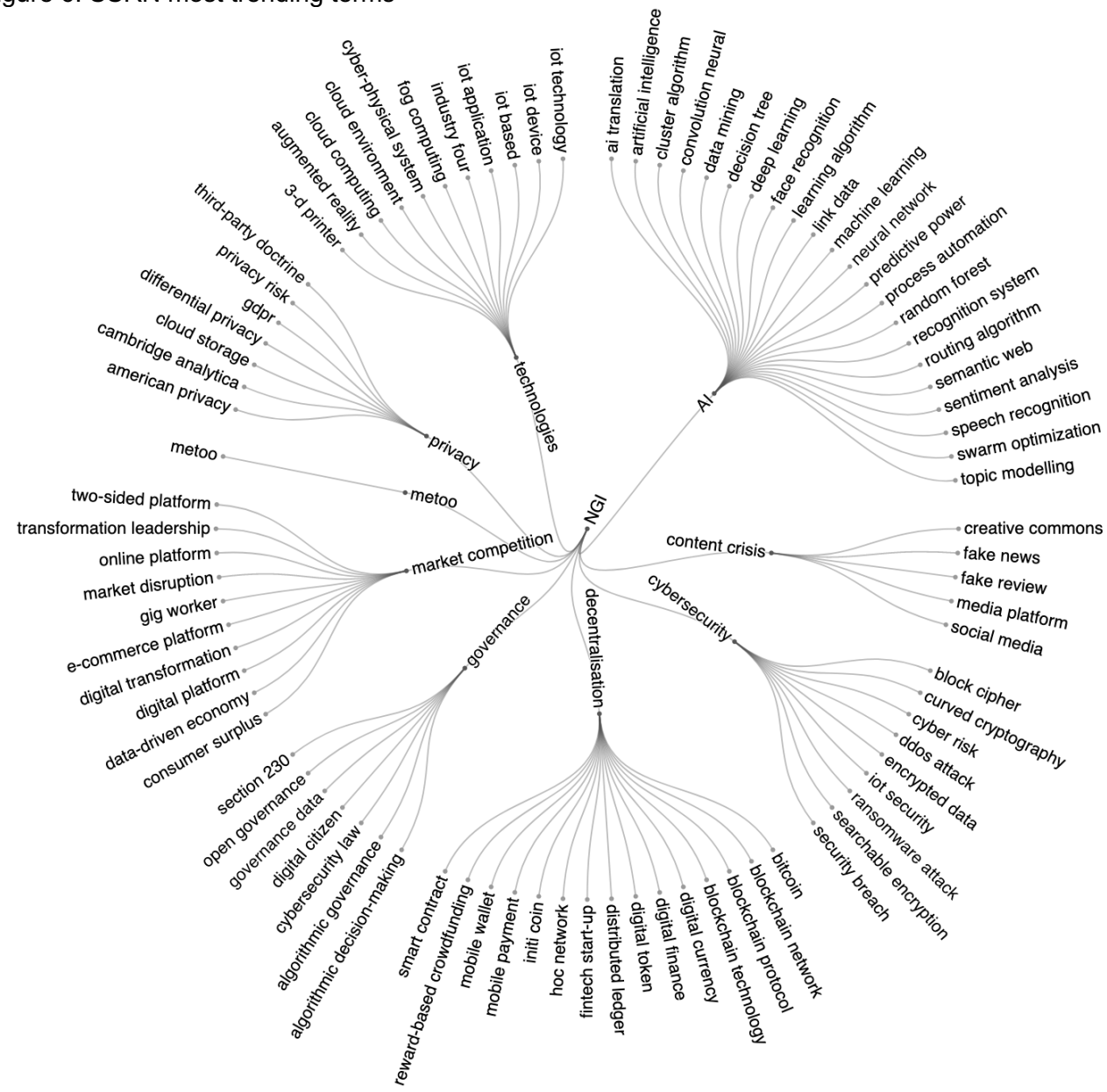
Figure 5. ArXiv most trending terms



For a 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.

AI and ML methods has been strongly present in social sciences as well. Common topics with Arxiv include decentralised technologies and internet of things. Researchers have increased focus in the field of market competition (platforms, platform competition, digital transformation), the content crisis (fake news, media platforms etc), issues related to policy and governance (Section 230, GDPR) and privacy.

Figure 6. SSRN most trending terms



4.3 Term frequencies over time

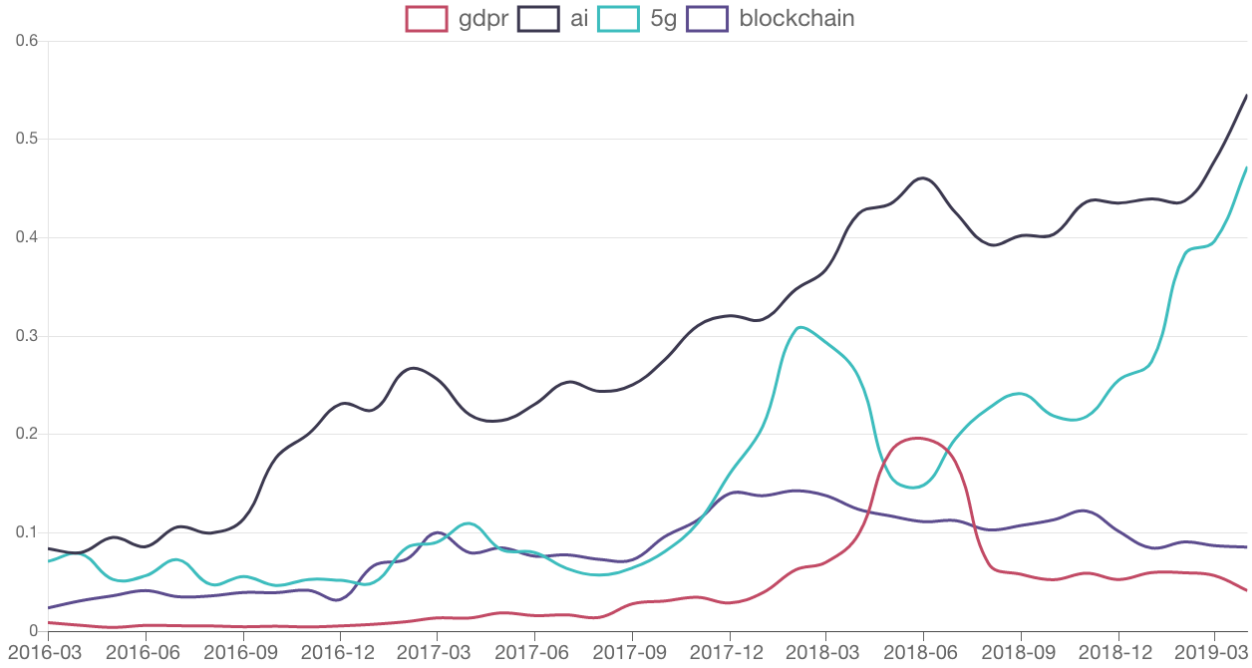
Next, the evolution of the frequency of trending terms can be examined across the three main sources: news media, STEM and social science working papers. The interactive application, available online (section: Trends), enables users to select any of the trending terms and track the average frequencies over time.

First, Figure 7. shows the evolution of the terms AI, GDPR, 5G and blockchain in online news sources.

AI has been the second most increasing unigram in the collected articles, showing a strong increase from the second half of 2016 until 2018. During 2018, the term frequency has been rather stagnating. In comparison, the term 5G has experienced periods of quick increase and decline. A similar seasonality can be observed in the case of GDPR: the frequency of GDPR

has been first gradually increasing, then skyrocketed around May 2018, when it came into force. After a short period of large media interest, the frequency declined.

Figure 7. Trending terms in news articles



In the case of social science working papers, the volatility of frequencies is very high, with strong increases and declines. In comparison, research on deep learning and reinforcement learning is steadily growing at Arxiv.

Figure 8. Trending terms in SSRN working papers

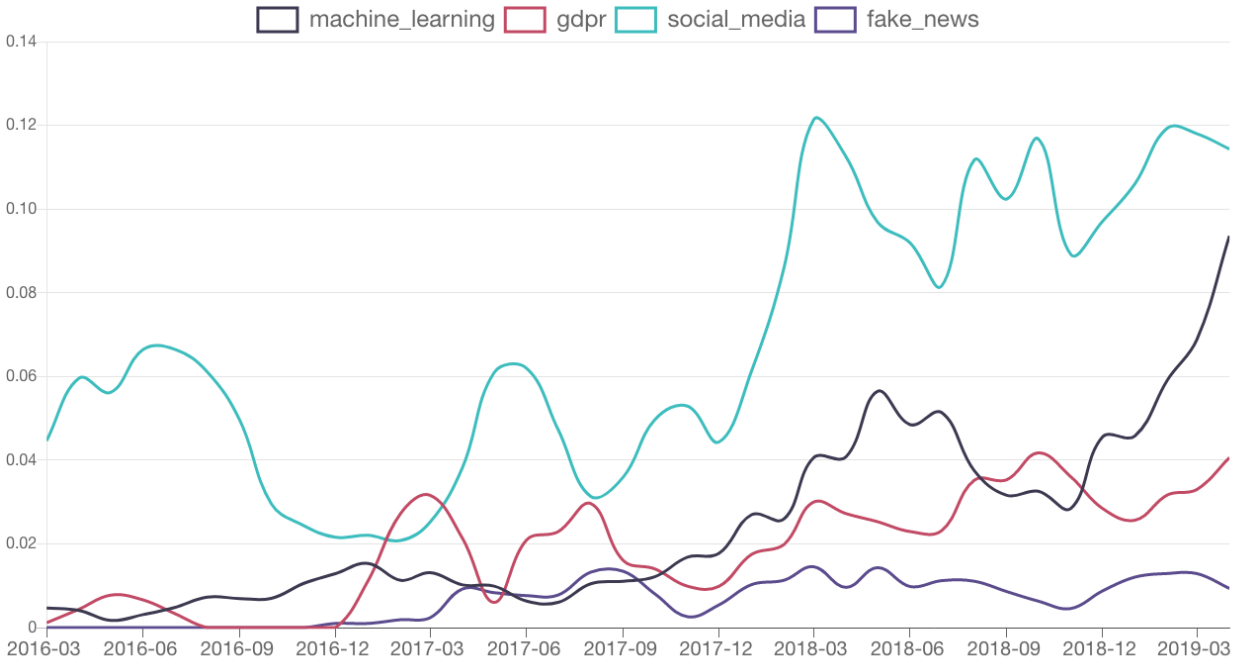
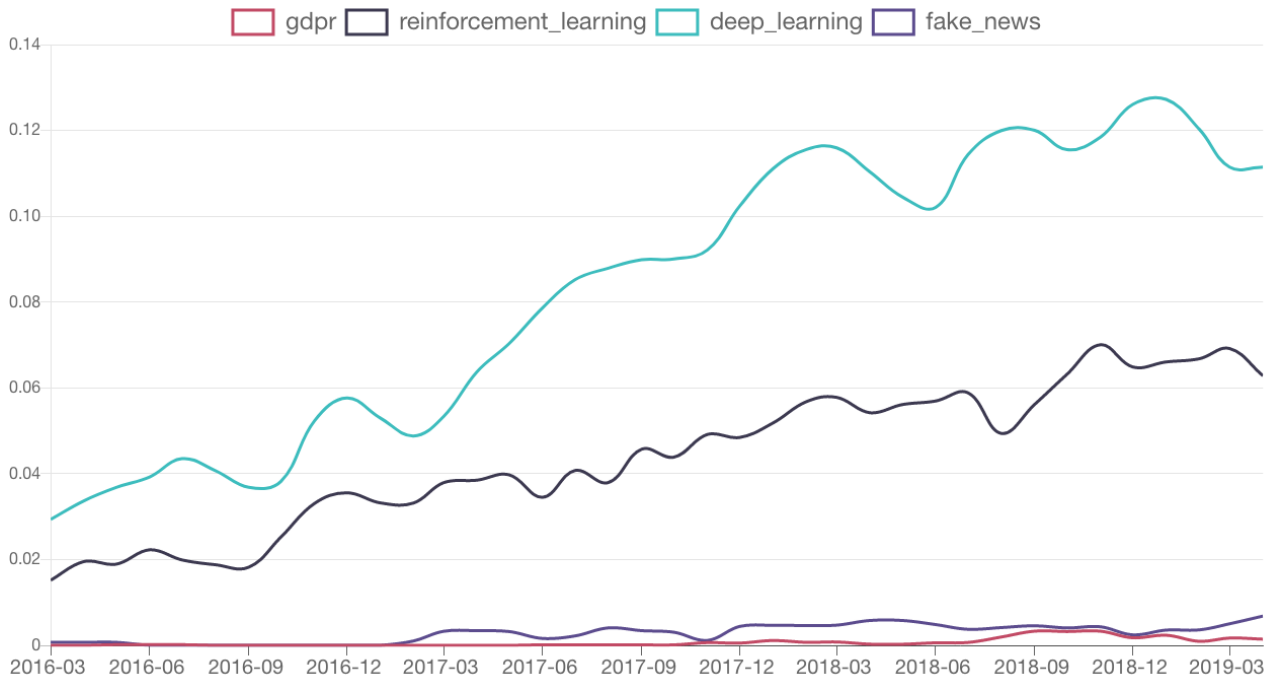


Figure 9. Trending terms in ArXiv working papers



Comparing the trends of the same bigrams across the three sources, important differences can be spotted. First, in the case of social issues, news media seem to provide earlier signals than academic research. The frequency of “fake news” massively increased at the end of 2016 (US presidential elections) in online news, while working papers began to cover this area 3-6 months later. This result is not surprising, given that online news media is reacting to daily events, while academic research has a much slower publishing cycle, even in the case of working papers.

Another interesting insight is related to the details and richness of the analysed texts. Comparing the trends of AI and reinforcement learning, the figures reveal that while the average occurrence of AI is much higher in online news, researchers are publishing on narrower fields, such as reinforcement learning. Therefore, the analysis of working papers provides a better overview of emerging technologies.

Figure 10. Frequency of the term “fake news” across the three sources

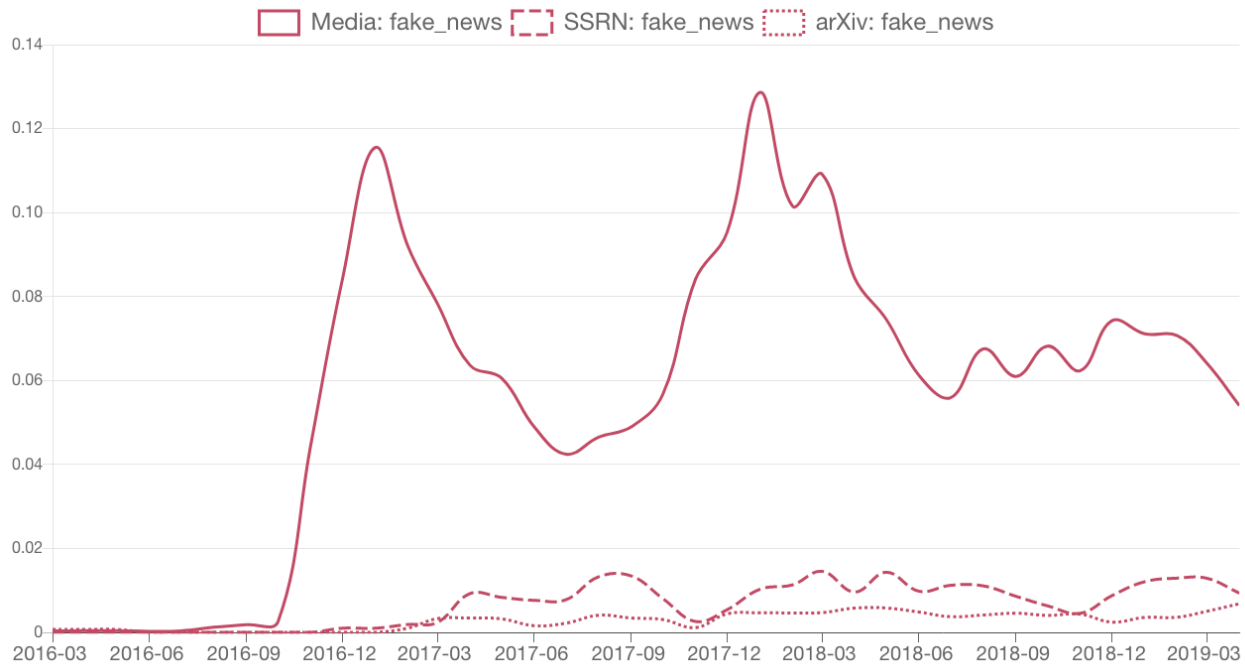


Figure 11. Frequency of the term “AI” across the three sources

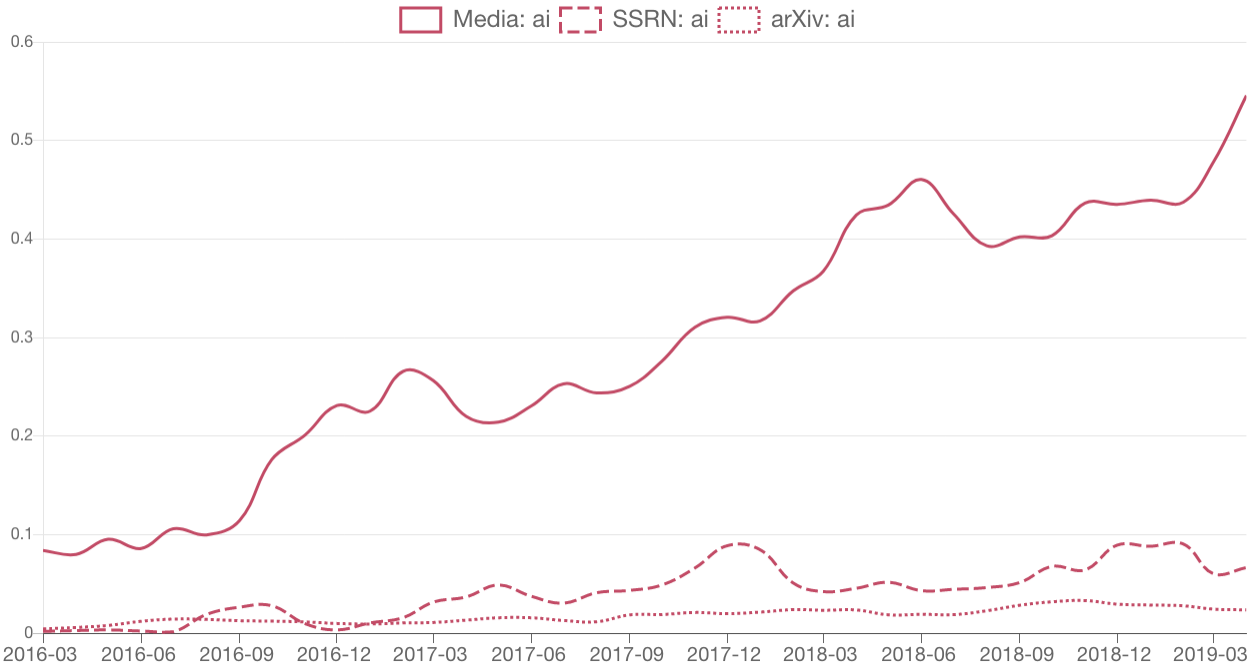
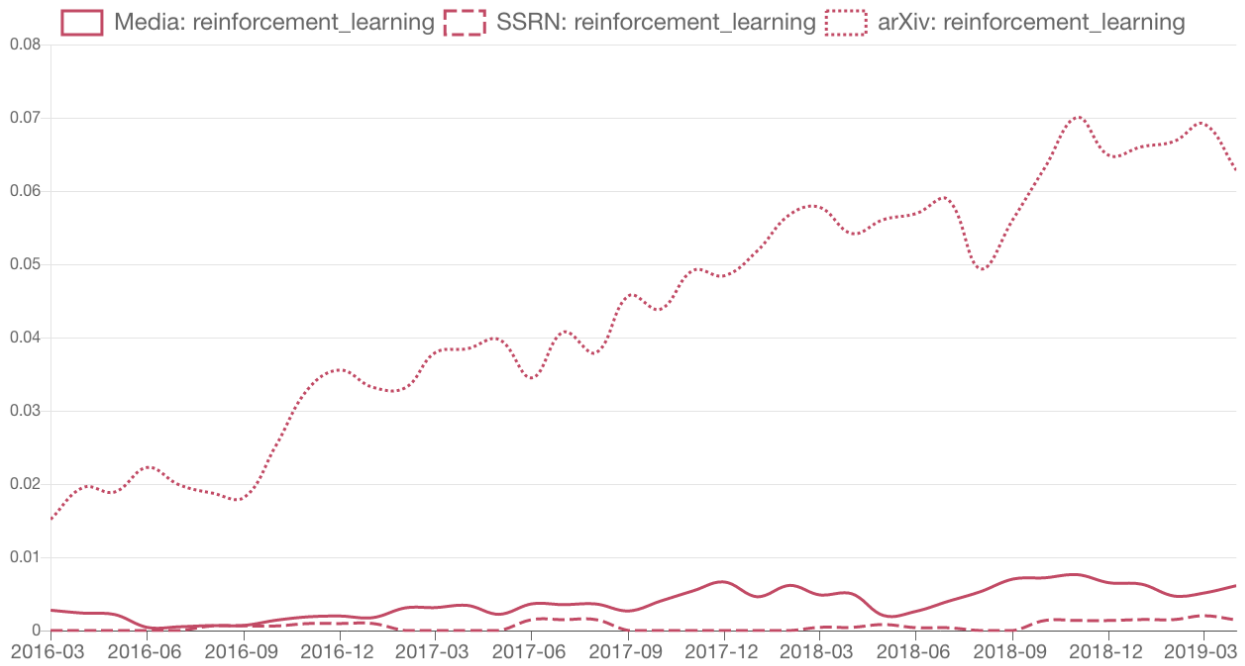


Figure 12. Frequency of the term “reinforcement learning” across the three sources



5 Deep dives

Following the identification of trending technologies and social issues, we continue the analysis with further exploration of selected umbrella topics.

Using co-occurrence and sentiment analysis, various details are extracted from news stories, including the relationship between trending terms, key persons and institutions.

For every deep dive topic, co-occurrences were examined for selected trending terms to reveal more details about news stories. In each figure, the first term provides the umbrella topic (e.g. AI), while the second column gives the selected terms for the analysis (e.g. project Maven). For each selected term, the frequently co-occurring words are presented.

In the case of the sentiment analysis, the perception of news stories is examined. The figures present the average monthly sentiment scores, also signalling the number of paragraphs used in the calculations (top 3 months by paragraph count). For each term, those co-occurring words are extracted in tables that appear in the paragraphs with the most positive and negative sentiments.

The results are additionally verified using topic modeling analysis. The LDA algorithm provides us the most important groups of terms in the analysed articles. For each topic of a deep dive, topic modeling is prepared for the subset of documents containing domain-specific terms. The list of terms used to filter the relevant articles is provided in the Appendix.

These methods were implemented for eight selected umbrella topics. The criteria for the selection of the topics are as follows:

- Does the trending technological area have established standards and regulations?
  Areas or technologies where specific EU-wide standards or regulations are yet to be developed and deployed are preferred. For example, while Roadmaps for 5G technology rollout are already in place, the development of regulation for blockchain or AI is at an early phase.

- Is the social issue already addressed at the EU level?
  We prioritise social challenges where policy or regulatory interventions have largely not been put in place. As an example, GDPR already provides a regulatory framework for privacy, while problems related to social media are currently debated.

- Are Internet technologies playing a crucial role in the social issue?
  We focus on those social and economic challenges where ICT technologies play a crucial role. Examples illustrating such challenges include the regulation of big tech monopolies or the expansion of the Chinese tech sector.

Therefore, we explore those topics in the deep dives, where social relevance is high, EU-level regulations are only partially in place, and the mapping exercise has significant value added.

Based on these criteria, the following 4 technological areas and 4 social challenges were chosen:

Technological areas	Social challenges
AI & ML	Social media & content crisis
IoT	Market competition
Blockchain & cryptocurrencies	Internet regulation
Quantum computing	Chinese tech

AI and ML

Artificial Intelligence and machine learning algorithms are among the most important computer science fields, with huge social implications. The top trending terms include both specific algorithms (e.g. reinforcement learning), tools (e.g. PyTorch) and also various controversial implementations, as deep fakes or Google's project Maven. Moreover, AI and ML may be crucial in solving many social challenges, as in the case of the content crisis on social media.

In the deep dive, we seek to answer the questions:

- What kind of risks are related to AI?
- Which social challenges may be solved by AI?

IoT

Internet of Things, along with various related technologies (AR/VR), has large potential to transform consumer electronics and production systems as well (industrial IoT). On the other hand, IoT devices raise cybersecurity and privacy concerns (e.g. smart speakers).

The deep dive will answer the questions:

- Which are the most promising implementations of IoT?
- Which are the greatest cybersecurity risks related to IoT?

Blockchain and cryptocurrencies

Blockchain has been long regarded as a transformative technology with large disruptive potential. Blockchain technologies may play a central role in the future of social media, financial services and in other intermediation services. As of today, the most widespread implementation of blockchain is related to cryptocurrencies. As an emerging technology, blockchain raises pressing regulatory issues.

We will answer the questions:

- Which are the use cases of blockchain?
- Which are the crucial regulatory challenges for blockchain and cryptocurrencies?

Quantum computing

Quantum computing, although there are promising developments, is not likely to become a mature technology in the next few years. However, quantum computing provides an opportunity for Europe to regain its competitive edge in advanced technologies. Therefore, mapping of quantum technology areas and developments has large value added.

Our research questions are:

- Who are the key players in the quantum field?

- What is the current state of quantum computing development?

Social media and content crisis

The spread of fake news, misinformation and the decline of trust in reliable sources create a profound challenge for the functioning of democracies and societies. While regulating platforms or implementing advanced topic filtering algorithms are among possible solutions, bringing back trust to written words may be far more complicated.

In the deep dive, we will answer the question:
- Which are the main threats for the digital public sphere?

Market competition

The giants of digital economy (GAFA: Google, Amazon, Facebook and Apple) are all functioning as platforms with incredible market power. While the US has been less active in regulating market competition, e.g. in the case of Facebook acquisition of rival Instagram and Whatsapp, the EU is leading the discussion on ensuring competition in the Digital Single Market.

The research questions are:
- What is the public perception of tech giants?
- What are the ideas for regulating the competition in the tech sector?

Internet regulation

Europe has been at the forefront of online regulations with GDPR, while the copyright directive (especially Article 11 and 13) has been more polarising among stakeholders. In the US, recent discussion has been focused on online content and Section 230 (platforms are not liable for the user generated content) or the controversial repeal of net neutrality rules.

In the deep dive, we seek to answer the questions:
- What are the negative externalities of the digital platforms expansion?
- Which Internet legislation attempts have caused heated public debates?

Chinese tech sector

China has managed to build a vibrant ecosystem in such key technologies as AI or 5G. The increasing position of the Chinese tech sector has brought a momentous challenge for both Europe and the US. China may be the forerunner in developing advanced AI systems and 5G networks, while advocating an approach to citizen rights and privacy that is in stark contrast to European values.

The research question is:
- What are the main concerns related to the expansion of Chinese tech sector?

5.1 AI and ML

The co-occurrences provide rich details on the social issues related to AI and ML algorithms. While AI can lead to new innovations and be helpful in various recent challenges (e.g. tackling fake news), these algorithms often work in a non-transparent way (“black box”), may be prone to biases, or implemented for questionable purposes (“killer robots”).

Algorithms have been in the centre of recent controversies, such as the Cambridge Analytica scandal, or the implementation of facial recognition at the Berlin Südkreuz station[5]. Another example is project Maven, a cooperation between Pentagon and Google to implement AI algorithms for the identification of people on drone footage[6]. The involvement of Google in the military usage of AI stirred intensive debate, including the protest of Google employees. The backlash in the company led to Google's resignation from further cooperation with the Department of Defence, e.g. in the JEDI project (Joint Enterprise Defense Infrastructure)[7]. Therefore, the ethical usage of AI is a key point of public debate. The co-occurrences enable us to identify crucial institutions (Pentagon, Google's Advanced Technology External Advisory Council), persons (academics Jonathan Zittrain and Joanna Bryson) and companies (Byton - Chinese electric car producer, AI start-ups Doxel, Clarifai etc) as well.

Facial recognition is used as a case study for the sentiment analysis. At the beginning of the explored time period, the articles on facial recognition were initially rather positive (compound score: 0.22 on a scale of -1 to 1; "voice assistant" and "AI technology" being the most positive connotations), with a significant decline at the end of 2017, possibly due to the increase in events reporting on the questionable usage of the technology.

On the one hand, this technology can be seen as a convenient tool for tagging photos in the social media or authorising mobile payments in a secure way. On the other hand, we observe growing privacy concerns around facial recognition applications in the marketing industry and law enforcement[8].

Opponents of the implementation of facial recognition argue that it is a fundamentally biased technology, *exacerbating existing inequalities in the criminal justice system, relying on databases and algorithms built on a history of discriminatory policing*[9]. Moreover, experts warn that facial recognition algorithms are misidentifying black people, women and young people at higher rates than older white men[10].

The topic modeling also provides insight into the areas where AI is being used: robotics, home IoT devices and online services. Autonomous cobots (collaborative robots), capable of performing independent decision making and action, while working in tandem with people, hold great promise in changing many areas of our lives.

Coming to consumer use, AI algorithms are the basis of personal voice assistants in smart devices, and are also increasingly used by online platforms, e.g. in content recommendation.

We also observe a lot of media interest about the impact of AI on social media, especially in the field of tackling fake news.

Conclusions: AI is the most trending area both in technology news, as well as in computer science research. AI may be useful in solving such challenges as fake news and content moderation. While it is already widely used in consumer products, services and in the industry, major risks are related to AI algorithms. The usage of AI in many cases violates privacy or produces biased results. Moreover, AI needs to be used for ethical goals.

---

[5] https://www.politico.eu/article/berlin-big-brother-state-surveillance-facial-recognition-technology/
[6] https://www.theverge.com/2019/2/4/18211155/google-microworkers-maven-ai-train-pentagon-pay-salary
[7] https://www.zdnet.com/article/google-heres-why-were-pulling-out-of-pentagons-10bn-jedi-cloud-race/
[8] https://www.theguardian.com/technology/2019/feb/15/how-taylor-swift-showed-us-the-scary-future-of-facial-recognition
[9] https://www.theguardian.com/technology/2018/jun/05/facial-recognition-us-mexico-border-crossing
[10] https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Figure 13. Co-occurrence analysis for AI and ML

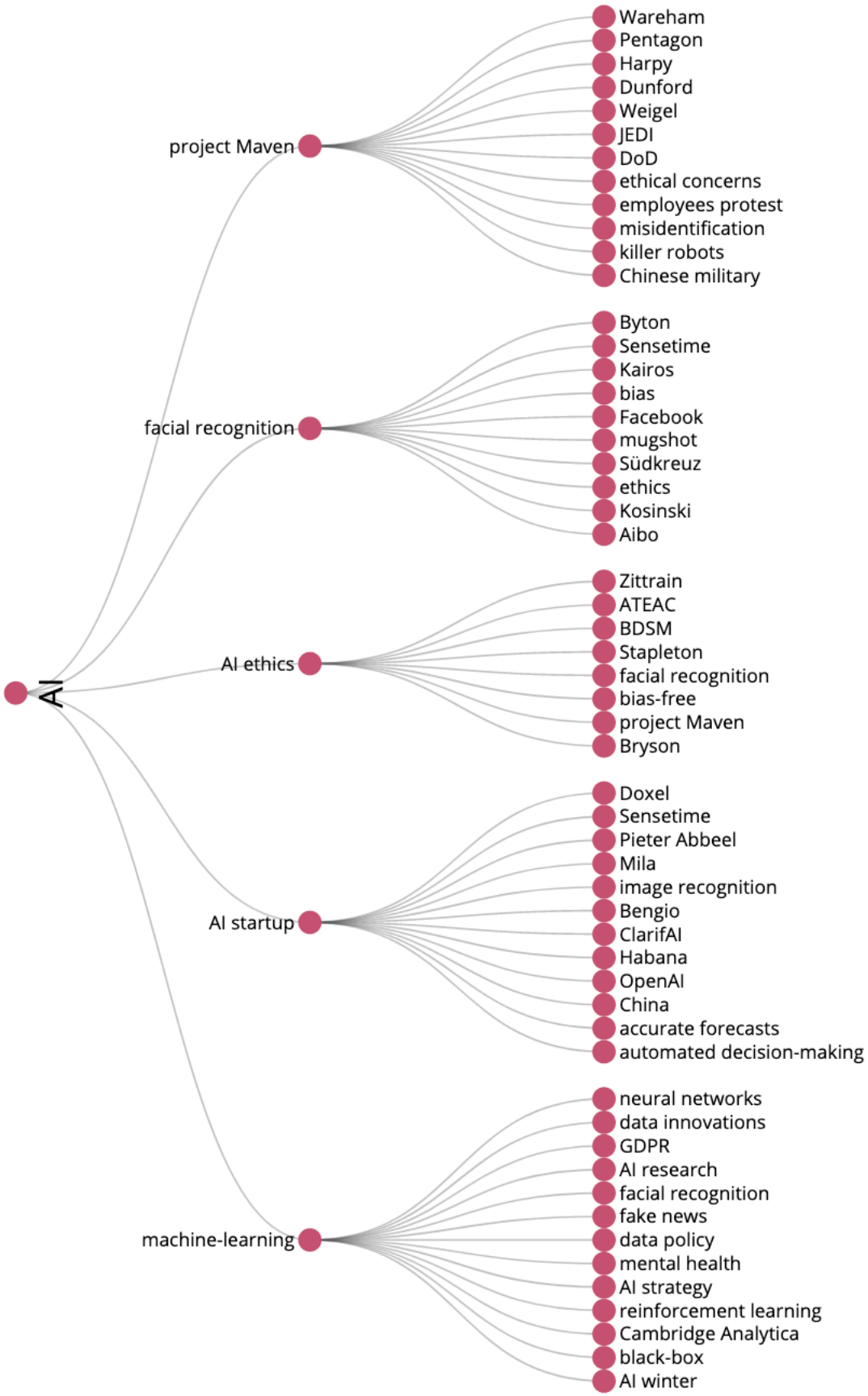


Figure 14. Sentiments analysis: facial recognition

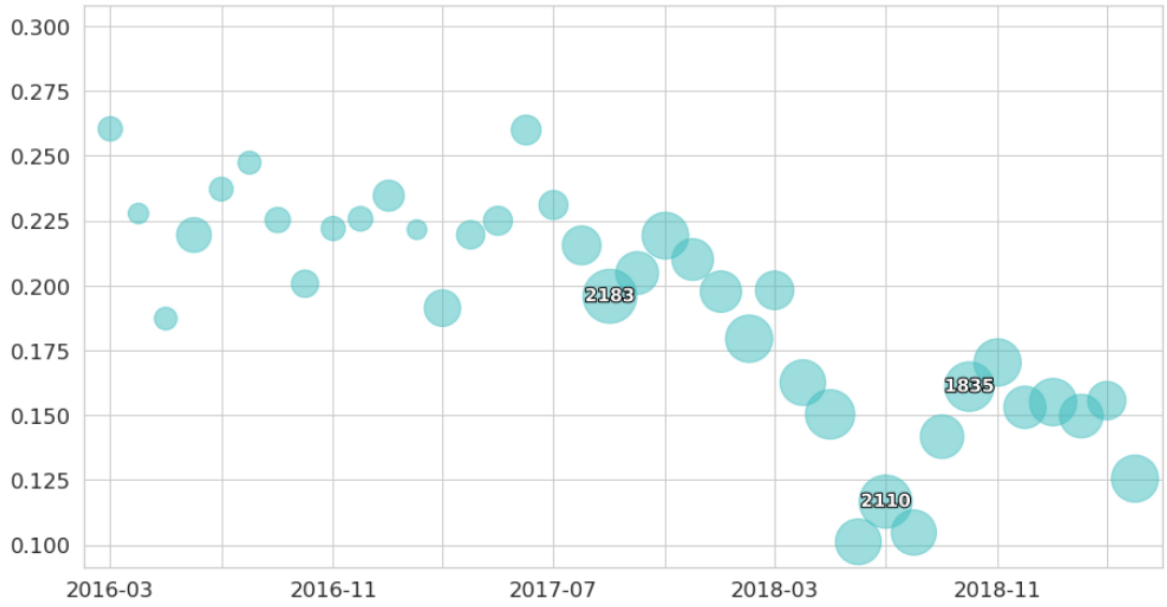


Table 1. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
voice assistant ai technology ai research edge computing ai startup	border guards autonomous weapons project maven big brother racial bias

Table 2. Topic modeling: AI and ML

Topic 1 Robotics (30.8% of tokens)	Topic 3 Social media regulation (17.2% of tokens)	Topic 5 Digital services (5.1% of tokens)
robot human people learning data technology researchers machine world computer	facebook data twitter people government social privacy media content law	uber protection act personal music apple amazon spotify service streaming
Topic 10 Home IoT (1.3% of tokens)	**Topic 11 Startups ecosystem (1.2% of tokens)**	**Topic 12 Fake news on social media (1.2% of tokens)**
alexa amazon echo home smart voice speaker alphago assistant google	disrupt battlefield startup tickets techcrunch alley storage berlin event judges	news facebook feed content publishers twitter stories fake trending algorithm

5.2 IoT

Three main aspects of IoT have been discussed in the tech media: industrial applications (IIoT), customer devices (e.g. smart speakers) and IoT security challenges.

In the field of industrial IoT, the German car making industry is a prominent player. The objectives of the Industrie 4.0 strategy is bringing connectivity for smart factories, machines and management systems.

The abundance of sensitive data produced by IoT sensors create challenges for the current model of co-location centres. Future development of IoT might require rethinking where companies place their data centers[11].

The Mirai botnet attacks, that took advantage of insecure network devices by attempting to log in using default passwords, raised concerns about the security of the IoT systems[12]. Some

[11] https://www.networkcomputing.com/data-centers/how-iot-will-redefine-colocation

[12] https://arstechnica.com/information-technology/2019/04/new-variants-of-mirai-botnet-detected-targeting-more-iot-devices/

solutions have been offered by security experts, including technologies being in the early development phase, such as quantum cryptography.

In the context of consumer IoT, the impact of e-Privacy regulation was discussed. The regulation is going to expand the definition of electronic communication services and protect data exchanges, involving situations in which non-personal data is being transferred, such as when two machines exchange data in M2M communications[13].

An important role of ML algorithms (e.g. neural networks, reinforcement learning) in the development of such IoT applications as voice assistants, e.g. to better handle not only different languages, but also their regional variants[14].

Sentiments around IoT in the tech media have been steady and positive in the analysed period. Our analysis suggests that industrial application of IoT and edge computing technology enabling advanced on-device processing are among the most positive associations with IoT. On the other hand, we observe security concerns related to IoT (e.g. Mirai botnet attacks). Furthermore, 5G networks necessary for mature IoT applications (e.g. increased M2M communication) are still in early development or under scrutiny after the Huawei scandals.

This polarisation is also reflected in the topic modeling analysis. The topics reveal that IoT has a huge impact on businesses (Topic 1) with various available enterprise platforms (Topic 2- e.g. Aruba, Gemalto etc). Moreover, anyone can build smart home solutions with simple IoT solutions (Topic 12- Raspberry Pi and Arduino). But its development highly depends on 5G networks (where Chinese Huawei seems to have an advantage) and is threatened by security flaws.

Conclusions: IoT is essential for future production systems, and also has a wide range of implementation for personal use. While some risks are similar as in the case of AI (they are often used together, e.g. for voice assistants in smart devices), the issue of cybersecurity is essential.

[13]https://www.euractiv.com/section/digital/opinion/eprivacy-is-about-the-entire-economy-and-we-need-to-get-it-right/

[14]https://techcrunch.com/2019/06/11/amazon-alexa-team-uses-machine-learning-to-better-handle-regional-language-differences/

Figure 15. Co-occurrence analysis for IoT

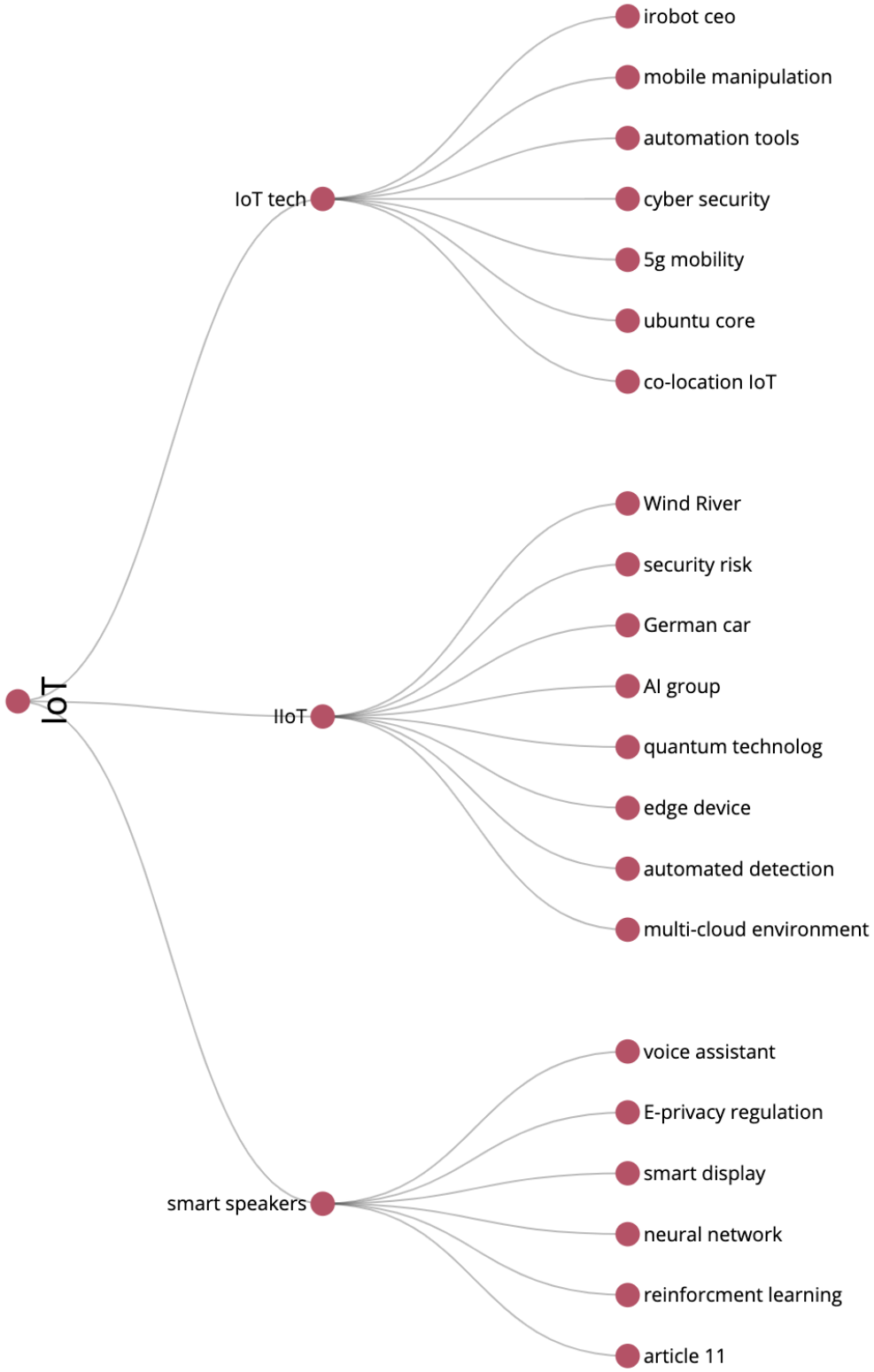


Figure 16. Sentiments analysis: IoT tech

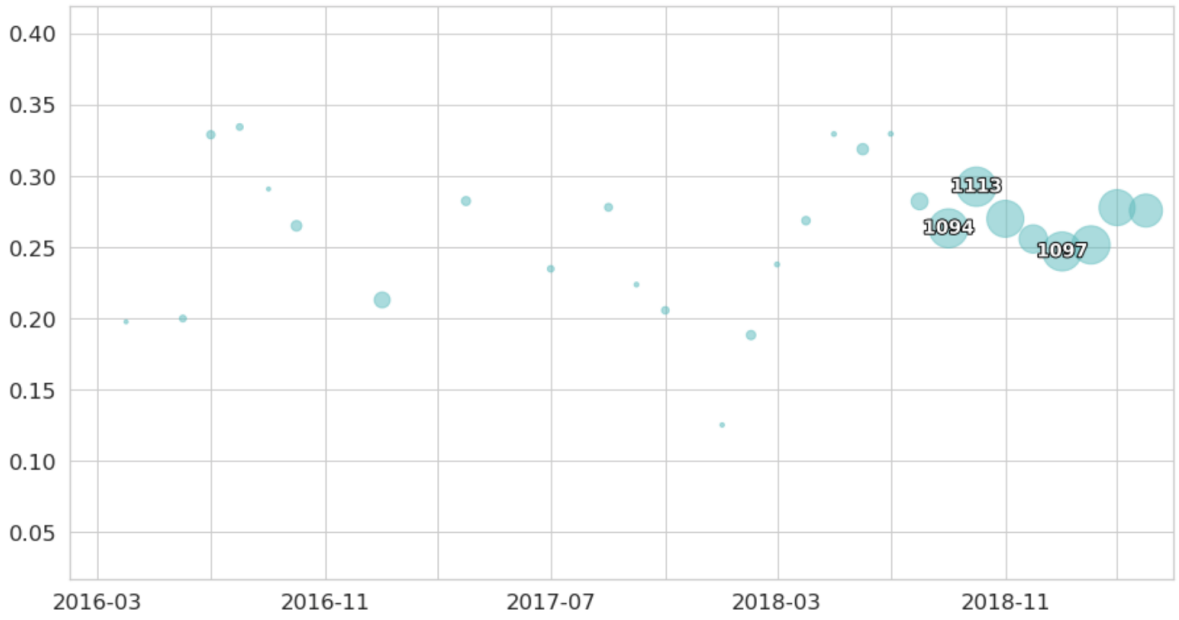


Table 3. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
industrial iot emerging technologies edge computing cloud vendors enterprise technology	security concerns security threats cyber security 5g network network equipment

Table 4. Topic modeling: IoT

Topic 1 Cloud (65.8% of tokens)	Topic 2 Cybersecurity (14.5% of tokens)	Topic 5 Telecoms (2.4% of tokens)
cloud	security	huawei
smart	attacks	network
services	google	nb-iot
business	device	australia
companies	privacy	spectrum
platforms	hackers	vodafone
network	vulnerabilities	telecom
software	users	cisco
connected	malware	korea
customers	android	optus
Topic 9 Botnets (1.1% of tokens)	**Topic 11 IoT platforms (0.7% of tokens)**	**Topic 12 IoT hardware (0.7% of tokens)**
botnet	bgp	raspberry
mirai	artik	arduino
ddos	aruba	board
malware	gemalto	suse
attack	aviv	glas
routers	dash	esim
infected	withings	thermostat
default	routing	raspbian
researchers	river	7nm
telnet	sds	euv

5.3 Blockchain and cryptocurrencies

The co-occurrences reveal the various potential areas where blockchain technologies can be implemented. One example for the usage of blockchain is the tracking of elements in the supply chain, e.g. in the food industry (Bumble Bee Food). The implementation of blockchain may be essential for solving the problem of post-Brexit Irish borders[15]. Similarly to physical goods, blockchain can be also used to control copyrights and track the ownership of intellectual property[16]. Blockchain is also a potential technology for secure social media platforms, prone to such data breaches, as in the case of the Cambridge Analytica scandal[17].

The co-occurrences confirm that blockchain is strongly connected to digital tokens and cryptocurrencies. Various start-ups have engaged in ICOs (Initial Coin Offering - the release of coins to collect funding, similar to Initial Public Offerings), e.g. in the area of trading with renewable energy[18]. While cryptocurrencies and digital tokens may have a huge potential in

[15] https://www.reuters.com/article/us-britain-eu-hammond-border/blockchain-may-resolve-irish-border-brexit-problem-hammond-idUSKCN1MB3FM
[16] https://www.zdnet.com/article/sony-explores-the-blockchain-to-create-drm-intellectual-property-protection-tech/
[17] https://www.theguardian.com/commentisfree/2018/mar/21/blockchain-privacy-data-protection-cambridge-analytica
[18] https://techcrunch.com/2018/05/08/green-power-exchange-enables-peer-to-peer-energy-sharing/

peer-to-peer services, the co-occurring terms also show the crucial regulatory challenge of the technology (e.g. Ponzi Scheme, potential risk). Cryptocurrencies are also used by criminals and hate groups to hide their financial activities[19].

Finally, cryptocurrencies are also examined by state institutions (e.g. UK Treasury Committee). Recent news include the Venezuelan attempt to mitigate the economic crisis by releasing a state-backed cryptocurrency, Petro[20].

We observe volatile sentiments around blockchain technologies in the tech media. It is most probably related to its slower than expected roll out and problems with building practical blockchain-based business applications. The period of rapid price surge of blockchain based cryptocurrencies in 2017 was followed by a trough of disillusionment in 2018. There have also been recurring opinions saying that blockchain based cryptocurrencies share many characteristics with the Ponzi scheme[21]. The decline of sentiment around blockchain was fueled also by scandals related to embedding illegal content in the blockchain network like child abuse images[22].

Topic modeling showed that media narratives on blockchain besides tracking more or less successful ICOs were focused on the hardware aspect of its development. The rapid proliferation of blockchain based cryptocurrencies has benefited companies manufacturing GPUs well suited for cryptomining.

Conclusions: Blockchain enthusiasm has tempered. In order to revive it, blockchain practitioners need to build more practical business and social applications reaching beyond cryptocurrencies. The potential areas where blockchain can be transformative include supply chain management, copyright control, social media and peer-to-peer transactions. However, blockchain requires additional scrutiny, as cryptocurrencies can be used for money laundering, and blockchain can also serve as a medium for illegal content.

[19] https://www.theguardian.com/commentisfree/2018/jan/24/bitcoin-currency-far-right-neo-nazis-cryptocurrencies
[20] https://www.zdnet.com/article/venezuelas-cryptocurrency-will-be-above-the-dollar-in-value/
[21] https://impakter.com/bitcoin-first-digital-ponzi-scheme-history/
[22] https://www.bbc.com/news/technology-47130268

Figure 17. Co-occurrence analysis for blockchain and cryptocurrencies

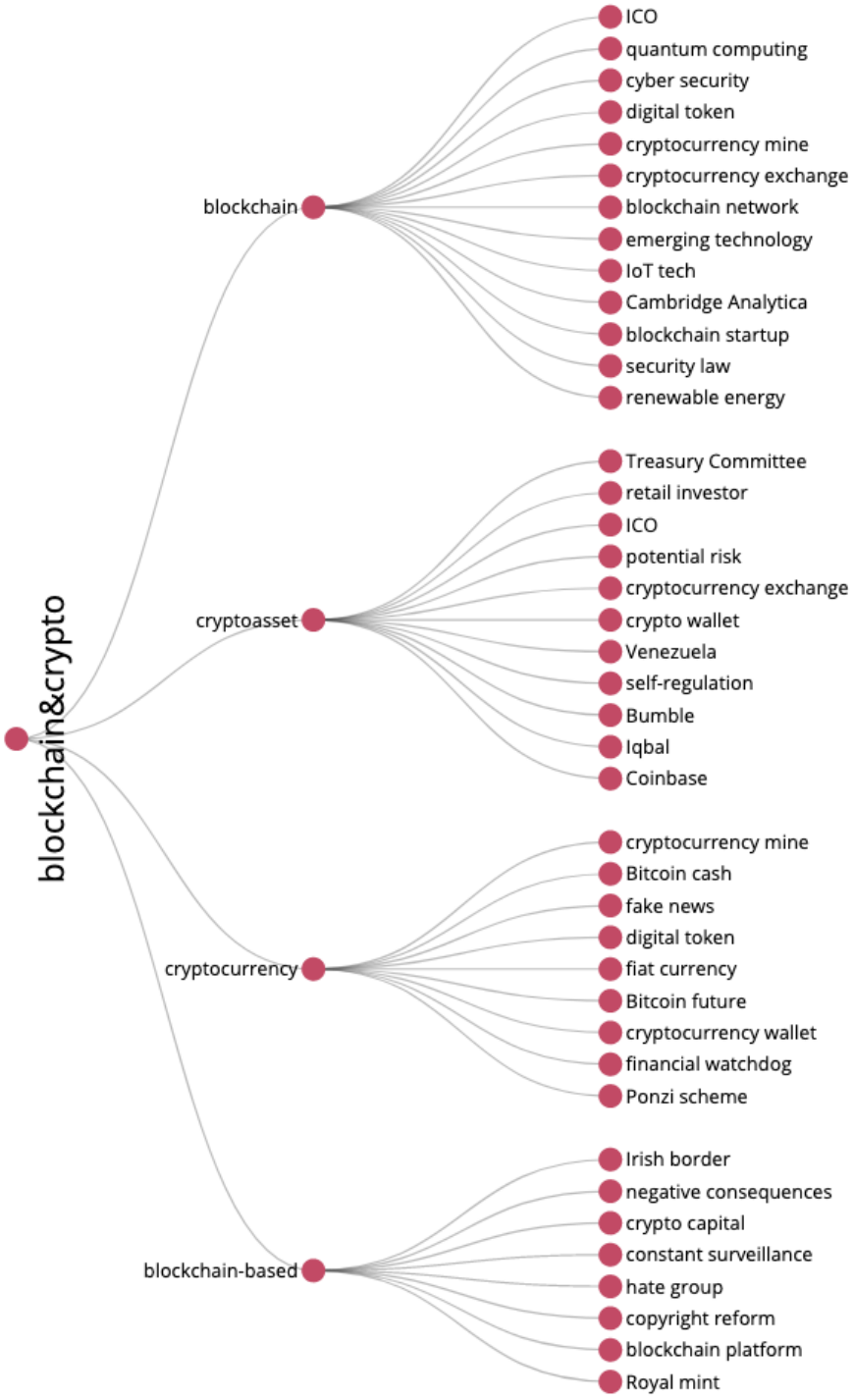


Figure 18. Sentiments analysis: blockchain

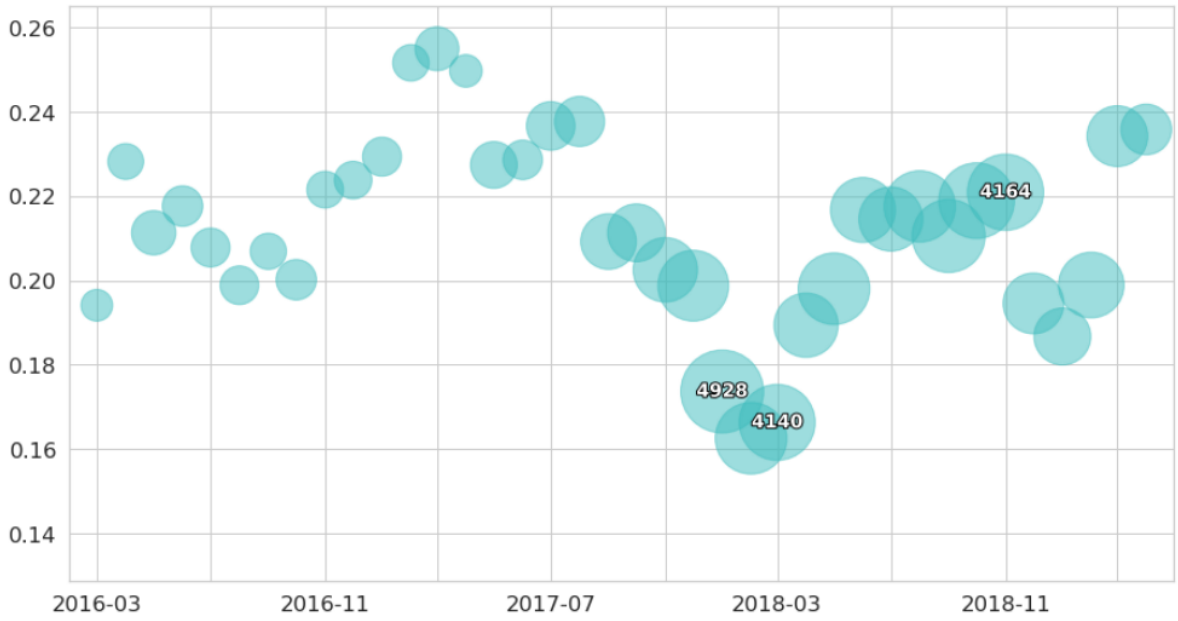


Table 5. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
renewable energy enterprise technology edge computing startup battlefield decentralized apps	child abuse alex jones cryptocurrency wallets conspiracy theories ponzi scheme

Table 6. Topic modeling: Blockchain

Topic 1 Cryptocurrencies (48.6% of tokens)	Topic 2 Cloud & IoT (30.1% of tokens)	Topic 3 ICO (8.4% of tokens)
bitcoin cryptocurrency digital technology financial users money companies people coinbase	data cloud iot expo technology google security tech mobile devices	data ico sec bitcoin facebook securities protection breach trading law
Topic 4 Cybersecurity (2% of tokens)	**Topic 7 Venezuela (1% of tokens)**	**Topic 8 Hardware (0.9% of tokens)**
malware mining monero code attack researchers windows malicious security coinhive	petro venezuela tezos maduro sanctions alphabay tusk oil low-code installment	nvidia gpus amd chips graphics revenue mining parity bitmain chip

5.4 Quantum computing

Quantum computing has the disruptive potential that may impact many different sectors, ranging from medicine to cryptography. However, this technology is far from ready to be commercially deployed in spite of what companies like IBM might claim. The 2019 IBM's Q System One, advertised as the first commercial quantum computer, should be treated rather as a symbol than a breakthrough, useful for further research on quantum computing itself, and not for typical use cases[23].

Quantum volume, the quantum computing performance metric determined by the number of qubits (basic unit of quantum information), is increasing quickly. As an example, Q System One is powered with a 20-qubit processor, while its predecessor had only 8 qubits. However, we are still years away from Quantum Advantage, i.e. the point where quantum applications deliver significant advantages to classical computers. According to IBM officials, Quantum volume would need to double every year to reach Quantum Advantage within the next decade[24].

[23] https://www.theverge.com/2019/1/8/18171732/ibm-quantum-computer-20-qubit-q-system-one-ces-2019

[24] https://www.zdnet.com/article/ibm-hits-quantum-computing-milestone-may-see-quantum-advantage-in-2020s/

Another step in bringing quantum computing closer to the general public has been made by British company D-Wave that unleashed its quantum optimizer via an application programming interface[25].

Mindful of the future development of quantum computing, experts are already debating how to secure future digital systems by developing post-quantum cryptography. National strategies and intelligence alliances are vital in this sphere. In light of the Chinese espionage scandals in the telecommunication industry, some experts are calling for a joint intelligence action against future quantum hacking danger (see: co-occurrence Five Eyes).
The sentiments around quantum computing in tech media are volatile but rather positive. Negative connotations are related to the Trump administration and intelligence alliance (Five Eyes).

The topic modeling experiment helps us identify key players in the quantum computing industry (e.g. IBM, Microsoft, Tencent). Furthermore, we observe interest in cloud-based quantum computing, which may widen access to this technology, as well as advances in new designs using graphene nanostructures.

Conclusions: Quantum computing is not a mature technology yet, however, there are already available test applications. Key companies include IBM, Microsoft (US) and D-Wave (UK), while China has also heavily invested in research. The major issue related to quantum computing is reaching the Quantum Advantage, which can quickly lead to national cybersecurity crisis due to the obsolescence of encryption methods.

[25] https://arstechnica.com/science/2019/03/d-wave-2000q-hands-on-steep-learning-curve-for-quantum-computing/

Figure 19. Co-occurrence analysis for quantum computing

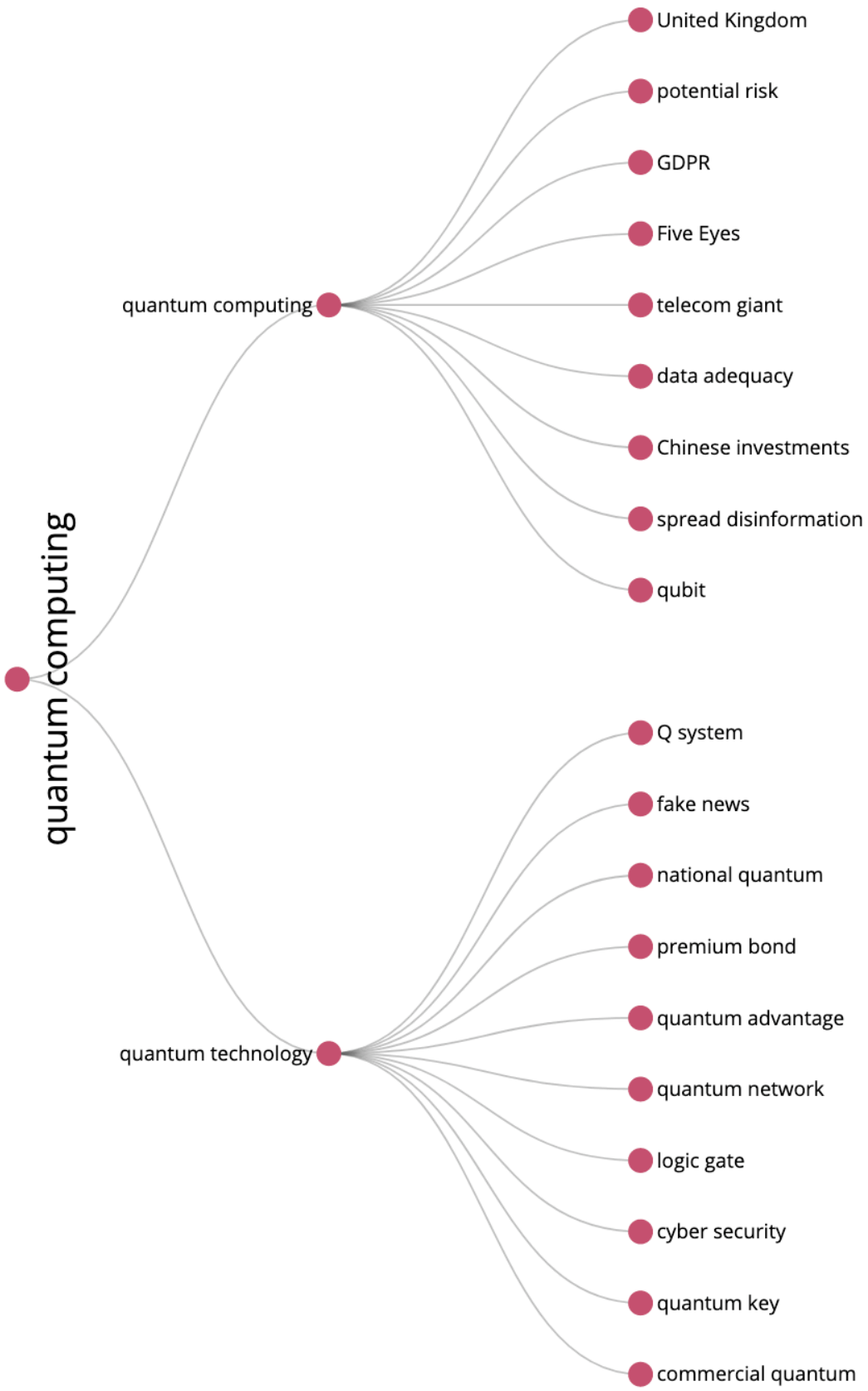


Figure 20. Sentiments analysis: facial recognition

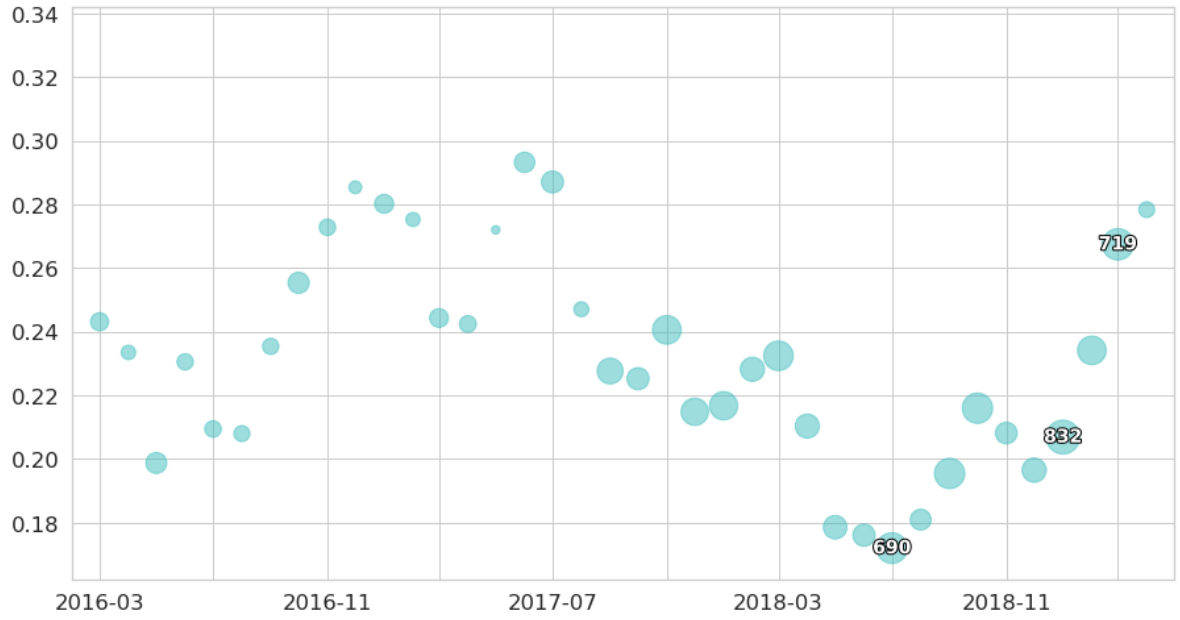


Table 7. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
reinforcement learning chinese researchers first demonstration information science nobel prize	logic qubit five eyes global trade trump administration quantum supremacy

Table 8. Topic modeling: Quantum computing

Topic 1 Qubits (45.4% of tokens)	Topic 2 Companies (35.8% of tokens)	Token 4 Graphene (2% of tokens)
qubit computer researchers light university atoms state physics energy work	data company ibm cloud computing microsoft security business network software	graphene nist transistor materials anu carbon voltage electrons dots cells
Topic 10 Cybersecurity (1.2% of tokens)	Topic 12 Quantum in the cloud (0.8% of tokens)	Topic 14 Quantum players (0.7% of tokens)
defense australian cyber qlabs government supercomputer quintessencelabs westpac probabilistic 49-qubit	aws blockchain vmware cloud ethereum slate multi-cloud cray kubernetes hyperledger	tencent zhang one-time bubbles pad apex holdings fee fears ted

5.5 Social media and content crisis

The spread of misinformation and fake news is among the key challenges for democracies. With the growing role of social media in providing information, sources that lack fact-checking and quality journalism are able to reach a wide audience. Moreover, social media platforms have been unable to tackle the misinformation spread by trolls and bots.

Fact-checking is often discussed in the context of fake news, deep fakes, and ranking algorithms that organise content at platforms. Recent developments include the flooding of Whatsapp with fake news, e.g. during the Brazilian elections[26]. The co-occurrences also present researchers of the field: Zittrain and Lyon.

The usage of conspiracy theories for political propaganda may not be a new tactic, but its efficiency has greatly increased due to social media. The results present the alt-right deep state conspiracy theory, spread by Alex Jones and his media outlet, Infowars.

Among “nation state” actors, Russia has been especially active in using social media to influence public opinion in various societies, including the EU and the US[27]. The report of

[26] https://techcrunch.com/2019/04/03/whatsapp-adds-a-new-privacy-setting-for-groups-in-another-effort-to-clamp-down-on-fake-news/
[27] https://www.theguardian.com/world/2019/jun/14/pro-kremlin-media-spread-false-eu-nazi-roots-european-elections

special counsel Robert Mueller provided evidence on systematic interference of Russia during the 2016 presidential elections[28]. The "13 Russians" refer to the indictment of 13 Russian citizens for tampering in the elections[29], while the operation is allegedly called projekt Lakhta[30].

Besides the problem of fake news, content moderation and censorship are crucial issues, especially in China (Google's Project Dragonfly).

Unsurprisingly, news stories covering conspiracy theories have rather negative sentiment. The most positive news (neutral sentiment) include Apple's "battery-gate" that was often mentioned in the context of planned obsolescence[31]. The story ended with a wider promotion to gain back consumer trust. Other neutral topics are related to the activities of platforms.

The most negative stories provide an overview of the greatest scandals, including conspiracy theories spread by the alt-right, and problems related to the Youtube Kids service. In the case of the former, Alex Jones is a key figure, as his Infowars spread various conspiracies on the mass shootings in Sandy Hook Elementary School[32]. Youtube Kids has reportedly exposed children not only to weird and disturbing content, but also to conspiracy theory videos[33].

The topic modeling also reflects the relevance of three problems: alt-right conspiracy theories, Russian meddling and Chinese censorship. Moreover, a cluster of articles reports on Wikileaks developments.

Conclusions: The Internet where fact-based arguments win over fake news seems to be a distant ideal. Online platforms seem to be unable to solve the problem of fake news, hate speech proliferation and foreign intelligence influence. Moreover, the Wikileaks revelations have posed important questions about the boundaries of free speech in the digital era.

[28] https://www.theguardian.com/us-news/2019/may/03/trump-putin-call-mueller-report
[29] https://www.euractiv.com/section/global-europe/news/purported-russian-hackers-stole-us-evidence-to-discredit-mueller-probe/
[30] https://www.zdnet.com/article/russian-national-charged-with-us-election-meddling/
[31] https://www.theverge.com/2017/12/20/16800058/apple-iphone-slow-fix-battery-life-capacity
[32] https://www.politico.eu/article/facebook-wades-deeper-into-censorship-debate-as-it-bans-dangerous-accounts/
[33] https://www.theverge.com/2018/4/6/17208532/youtube-kids-non-algorithmic-version-whitelisted-conspiracy-theories

Figure 21. Co-occurrence analysis for social media & content crisis

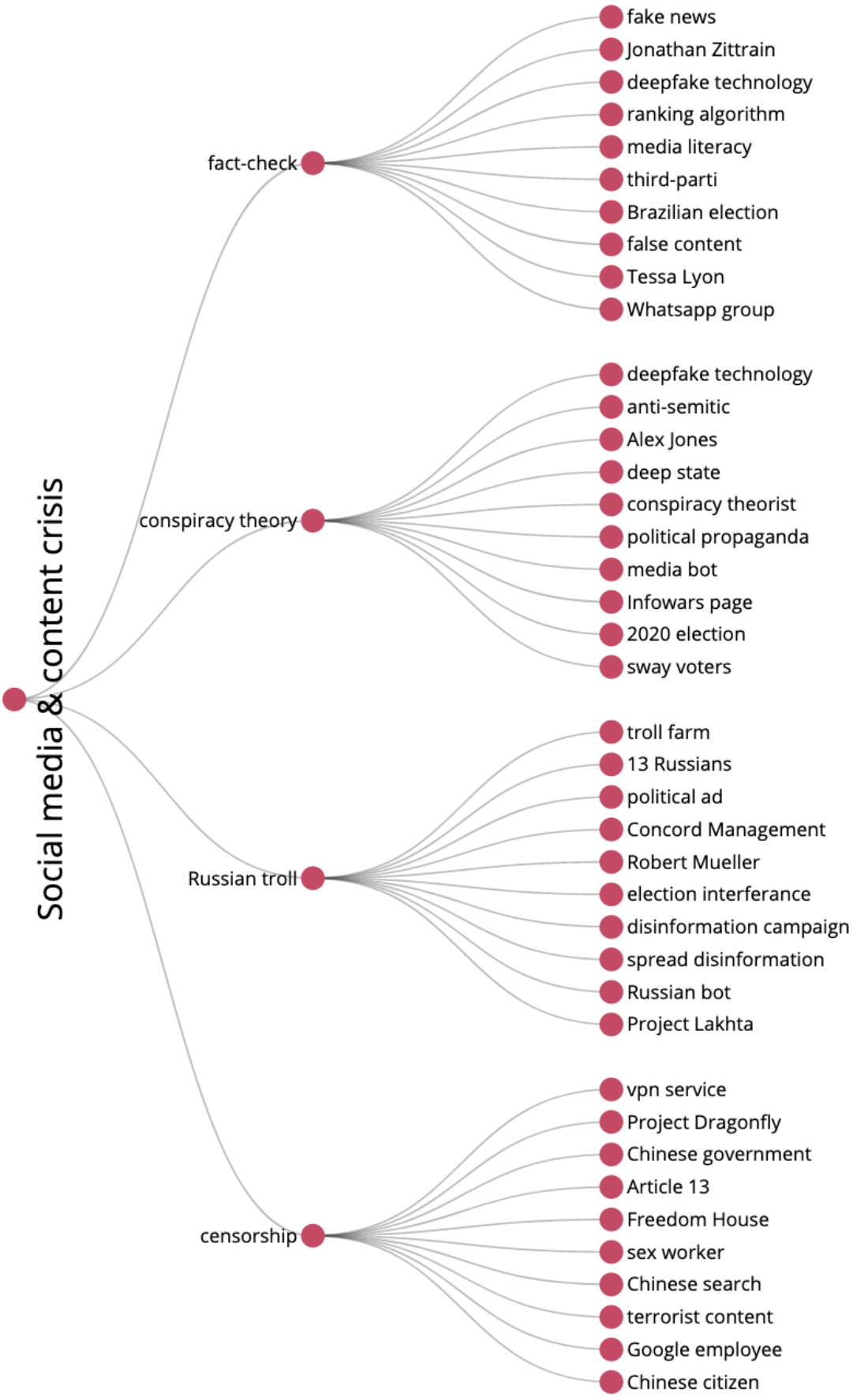


Figure 22. Sentiments analysis: conspiracy theory

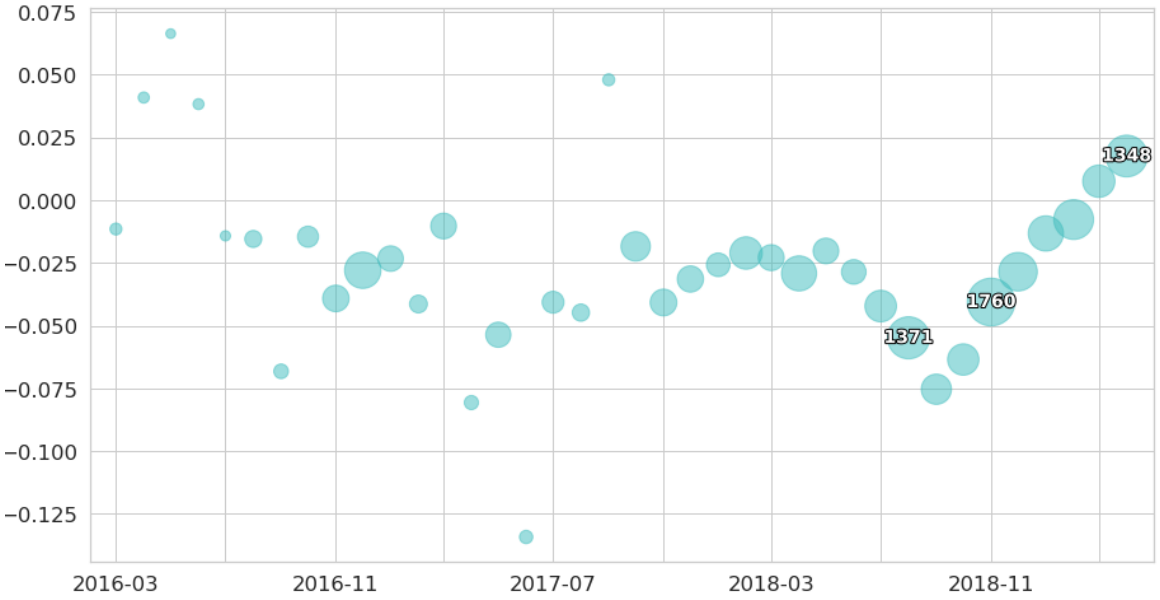


Table 9. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
cyber security battery replacement recommendation algorithms tech platforms cambridge analytica	mass shooting white supremacist sandy hook alex jones youtube kids

Table 10. Topic modeling: social media and content crisis

Topic 3 Streaming services and net neutrality (4.2% of tokens)	Topic 5 Alt-right (1.7% of tokens)	Topic 6 Russian hackers (1.6% of tokens)
netflix fcc streaming neutrality service amazon hulu cable verizon video	jones infowars influencer tay alt-right white yiannopoulos huffman charlottesville supremacists	russian fbi court charges police conspiracy guilt prosecutors fraud hacking
Topic 7 Chinese tech and freedom of speech (1% of tokens)	**Topic 15 Chinese tech and copyrights (0.5% of tokens)**	**Topic 24 Wikileaks (0.4% of tokens)**
china tencent alibaba wechat baidu tiktok beijing asia capital censorship	huawei copyrights piracy spectre infringement foxtel roadshow meng court intel	wikileaks assange cia hutchins equifax oettinger manning rpa fridge lynn

5.6 Market competition

"We need to address the power of digital platforms[34]."

-Competition Commissioner Margrethe Vestager, 03.06.2019

More and more politicians in the EU and the US argue that self-regulation has proven to be insufficient in the tech industry. The great concentration of market power created huge challenges: lackluster competition and monopolisation, blocking the entry of new firms by predatory pricing[35], underpaying workers (especially in "gig work", but not exclusively, e.g. "Stop Bezos" plan to tax Amazon for this reason[36]). There is increasing pressure for new legislation to reduce the problem of tech industry monopolisation[37].

[34] https://www.euractiv.com/section/digital/news/digital-brief-the-power-of-online-platforms/
[35] https://www.theverge.com/2019/5/13/18563379/amazon-predatory-pricing-antitrust-law
[36] https://www.theverge.com/2018/9/5/17819450/bernie-sanders-stop-bezos-amazon-worker-pay-corporate-welfare-tax-bill
[37] https://www.euractiv.com/section/digital/news/digital-brief-the-power-of-online-platforms/

The EU has been especially active in the field of antitrust and competition policy. Among the most important tech stories were the various fines against Google. The antitrust fine totalling 1.5 bln EUR was issued to the abusive practices against AdSense clients[38]. In 2018, Google was fined a record 4.3 bln EUR for bundling Android with Google Play Store, and in 2017 2.4 bln EUR for shopping search results manipulations[39].

On the other side of the Atlantic, there is also an intensive debate on antitrust. Key supporters of antitrust investigations include Senator Elizabeth Warren, calling to break up major tech firms[40], and Rep. David Cicilline, who is the chairman of an antitrust subcommittee[41]. On the other hand, the repeal of net neutrality aims at increasing competition and investments with deregulation at the telecommunication sector. Interestingly, FCC Chairman Ajit Pai also supports regulating major tech companies[42].

Another key challenge is related to the tax optimisation of tech giants. Recently, G20 financial ministers agreed to compile common rules to close loopholes used by global tech giants for tax evasion[43]. OECD also works on a broad coalition to introduce a digital tax. The program aims at setting a global minimum corporate tax and revamping countries' rights to tax foreign companies that have offshored their operations to low-tax areas[44]. There have been also support for an EU wide digital tax, initiated by Bruno Le Maire, French Minister of Economy and Finance[45].

The sentiment of stories on major tech giants have a downward trend since 2017, suggesting a growing concern over their social impact. Analysing positive and negative sentiments, the results show that while emerging technologies developed by these firms have positive scores, news stories reporting on social media platforms are significantly negative.

The topic modeling also suggests that there is a growing consensus among regulators on the need of regulating tech giants, and on the necessity to close loopholes of corporate taxes (Topic 2). Tech journalists showed also interest in highly competitive startup ecosystems and blockchain-based solutions in innovative financial solutions.

Conclusions: A worsening media image of tech giants can be observed. There is a growing consensus that previous soft antitrust approach has not been effective, and major tech companies have too great market power.

[38] http://europa.eu/rapid/press-release_IP-19-1770_en.htm
[39] https://www.theverge.com/2019/3/20/18270891/google-eu-antitrust-fine-adsense-advertising
[40] https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c
[41] https://www.zdnet.com/article/congress-opens-up-the-latest-tech-antitrust-front/
[42] https://www.theverge.com/2019/6/12/18662862/ajit-pai-fcc-facebook-google-amazon-apple-regulation-net-neutrality
[43] https://www.euractiv.com/section/digital/news/g20-agrees-to-wrap-up-big-tech-tax-rules-by-2020/
[44]https://www.euractiv.com/section/digital/news/global-roadmap-takes-step-toward-solving-digital-tax-conundrum/
[45] https://www.euractiv.com/section/digital/news/le-maire-campaigning-for-a-tax-on-tech-giants/

Figure 23. Co-occurrence analysis for market competition

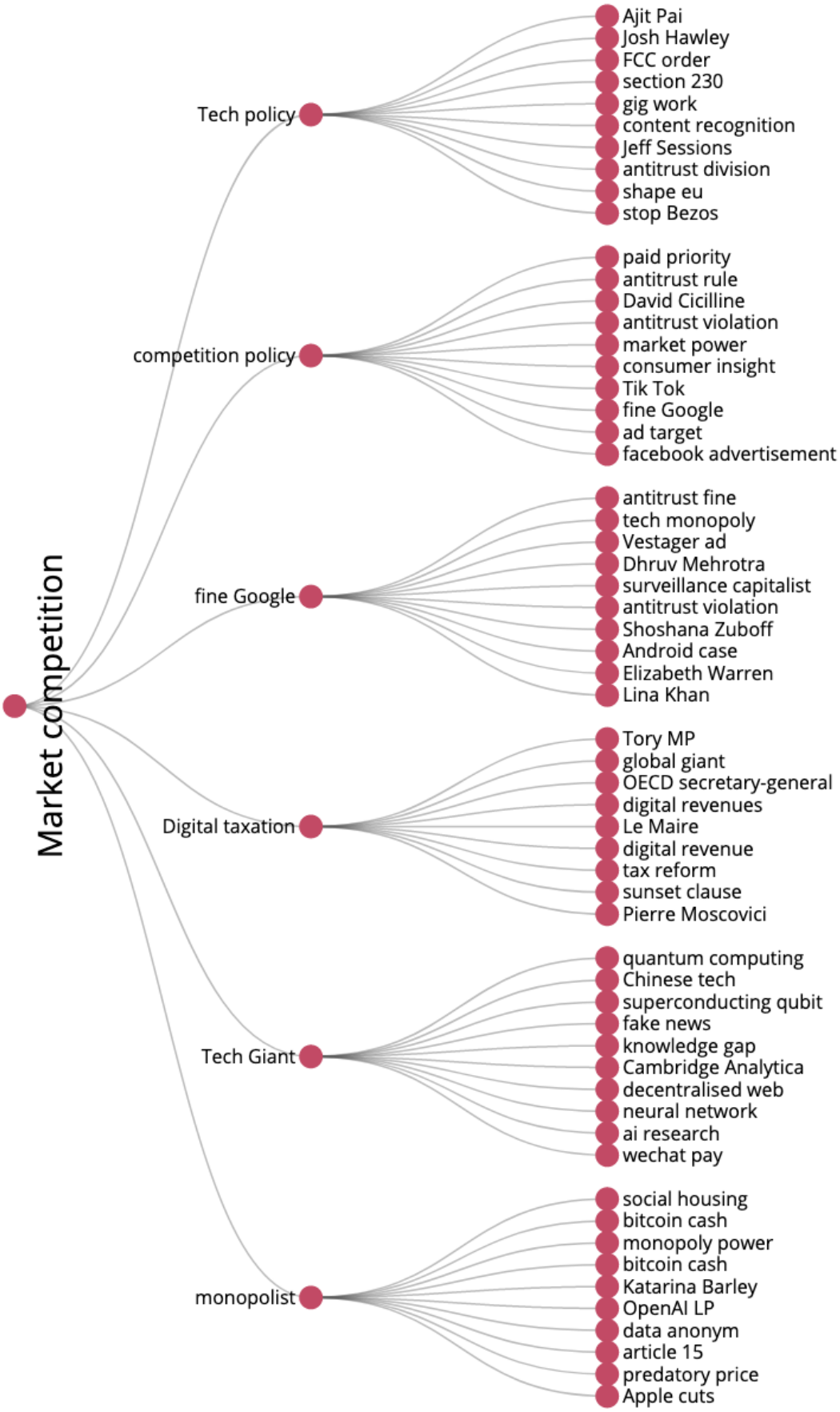


Figure 24. Sentiments analysis: tech giants

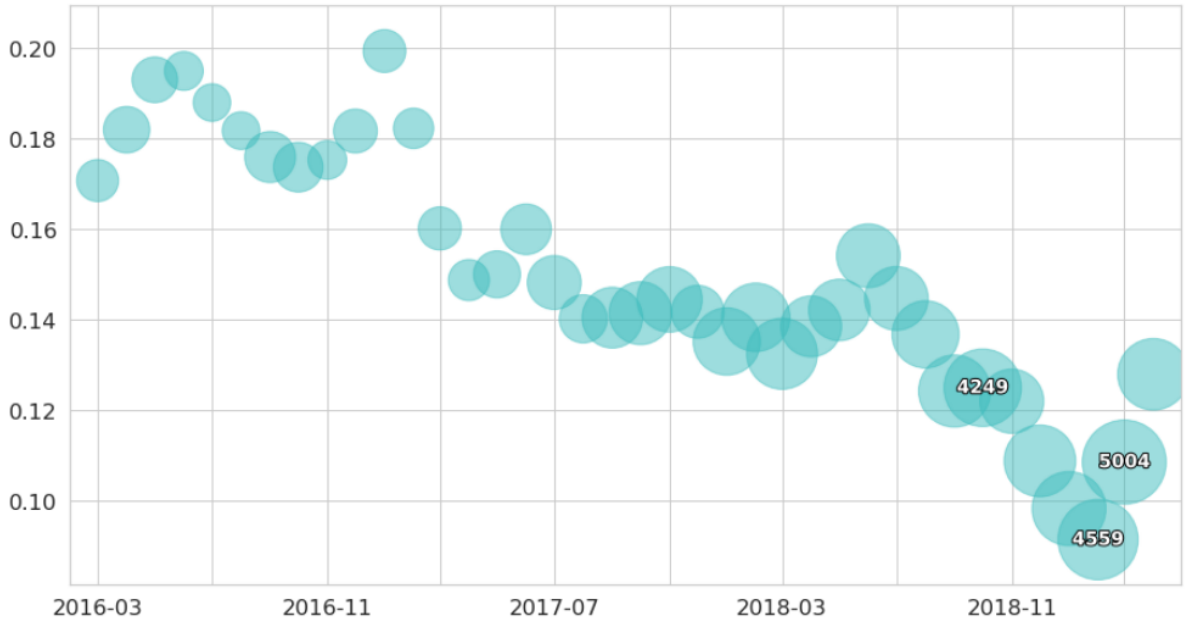


Table 11. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
renewable energy neural network ai systems voice assistant quantum computing	child sexual remove illegal alex jones terrorist content conspiracy theories

Table 12. Topic modeling: market competition

Topic 2 Digital tax (10% of tokens)	Topic 3 Startups (9.6% of tokens)	Topic 7 Blockchain (1.9% of tokens)
european security commission tax government data google digital huawei companies	startup capital funding investors ventures india platform robotics disrupt battlefield	blockchain bank payments bitcoin singapore asia financial airbnb fintech cryptocurrencies
Topic 12 Internet providers and net neutrality (1.2% of tokens)	**Topic 14 Streaming services (0.9% of tokens)**	**Topic 18 Smart devices (0.6% of tokens)**
fcc verizon t-mobile neutrality broadband sprint comcast pai cable wireless	music spotify snapchat youtube streaming instagram pandora app facebook artists	fitbit smartwatch fitness apple wearables idc pebble garmin heart health

5.7 Internet regulation

The next section covers recent regulatory issues both in the EU (GDPR, copyright directive) and in the US (net neutrality, section 230). In the last decade, major online platforms expanded with business models based on the provision of free services in exchange for personal data ('surveillance capitalism'). The largely unregulated development of platforms created numerous negative externalities: the loss of online privacy, filter bubbles, the spread of fake news or the decrease of control over creative content.

The General Data Protection Regulation (GDPR) has come into force in May 2018, providing a framework for handling the personal data of EU users. The regulation empowers users to control the usage of their personal data, object to the processing of personal data for marketing purposes, or request their personal data in appropriate machine-readable format (data portability). GDPR, besides setting the rules for the handling of personal data for companies, also requires the reporting of data breaches to supervisory authorities. Various major platforms did not fulfill these requirements and had been fined, including Google[46]. Data brokers, companies that aggregate personal data from various sources, had also been

[46] https://www.theguardian.com/technology/2019/jan/21/google-fined-record-44m-by-french-data-protection-watchdog

struggling to cope with GDPR[47]. Another recent news story include Amazon that following a data portability request sent Alexa voice samples to the wrong individual[48].

The EU Copyright Directives, published in May 2019, revised the copyright rules for the Digital Single Market. It has stirred heated debate in the context of major platforms like Google, especially for two articles: Article 11 ("the link tax") and Article 13 ("upload filter"). Article 11 requires platforms to pay licensing fee for creative content, while Article 13 requires them to prevent the upload of copyrighted materials by users[49]. The critics of the directive, including the world wide web creator Tim Berners-Lee, argued that the new rules may lead to censorship and the extensive implementation of content filters.[50] The co-occurrences also reveal the Directive's rapporteur, MEP Axel Voss.

Coming to more US-centric regulations, Section 230 plays a key role in the discussion on the responsibility of platforms in tackling such challenges as the spread of fake news or hate speech. Section 230 of the Communications Decency Act protects online platforms from liability of content published by their users. The rule has been designed so that platforms can moderate discussions in good faith and will not be held liable for their editorial decisions. However, for certain areas platforms are accountable, such as posts related to sex-trafficking, which is a result of the recently passed FOSTA-SESTA bill[51]. The responsibility of major platforms, such as Youtube, is widely debated for the spread of harmful or explicit materials. A recent example is the exposure of disturbing communities on Youtube by Matt Watson.

Finally, the US is polarised in the issue of net neutrality. Net neutrality guarantees that users can access online content and services without interference from Internet Service Providers, therefore banning such actions as throttling or paid prioritization. The net neutrality rules were adopted in 2015, and repealed in 2018. On the state level, California has been aiming at restoring net neutrality, conflicting with the federal government[52]. Some key actors of the news stories are presented in the co-occurrences: Ajit Pai (Federal Communications Commission Chairman) and Xavier Becerra (California's Attorney General).

The number of paragraphs show how the media extensively covered GDPR during its introduction (May 2018). The sentiment has been rather positive or neutral, with a significant decrease in the period following GDPR became effective. This decline may be related to the fact that news stories reporting on GDPR often cover data breaches, as in the case of British Airways[53]. Other concerns were expressed in relation to handling password data as plain text by Facebook[54] or managing health data[55].

The number of articles covering the copyright directive is much lower than in the case of GDPR, therefore it is difficult to assess changes in the trends. However, the positive and

[47] https://www.engadget.com/2018/11/08/gdpr-data-brokers-complaints/
[48] https://www.theverge.com/2018/12/20/18150531/amazon-alexa-voice-recordings-wrong-user-gdpr-privacy-ai
[49] https://www.theguardian.com/media/2019/mar/26/meps-approve-sweeping-changes-to-copyright-law-european-copyright-directive
[50] https://www.theguardian.com/technology/2018/jun/20/eu-votes-for-copyright-law-that-would-make-internet-a-tool-for-control
[51] https://www.theverge.com/2018/11/28/18115776/josh-hawley-section-230-ted-cruz-republicans-internet-liability
[52] https://www.theverge.com/2018/10/1/17922674/us-government-sues-california-over-net-neutrality-law
[53] https://www.theguardian.com/business/2018/sep/07/ba-says-hack-hit-only-those-buying-tickets-in-two-week-period
[54] https://www.zdnet.com/article/facebooks-latest-privacy-scandals-opens-regulator-floodgates/
[55] https://www.zdnet.com/article/conways-law-and-healthcare-data-management-genome-data-blockchain-and-gdpr/

negative co-occurrences well reflect the two sides of the debate. Paragraphs with most positive sentiment discussed European creators and publishers, while the negative ones covered the aspects of content filtering. Youtube and its CEO, Susan Wojcicki have been among the strongest opponents of the directive[56].

The topic modeling results show that internet regulation coverage in tech media were focused on the major legislations (GDPR, copyright reform, net neutrality), major digital platforms (ridesharing) and the free speech boundaries in the Wikileaks and NSA era.

Conclusions: The era of self-regulating tech industry seems to be over as legislators rush to regulate the Internet. Next years will show whether these efforts will shift power back to users, create more fair competition and reduce negative externalities.

[56] https://www.theverge.com/2018/11/12/18087250/youtube-ceo-copyright-directive-article-13-european-union

Figure 25. Co-occurrence analysis for Internet regulation

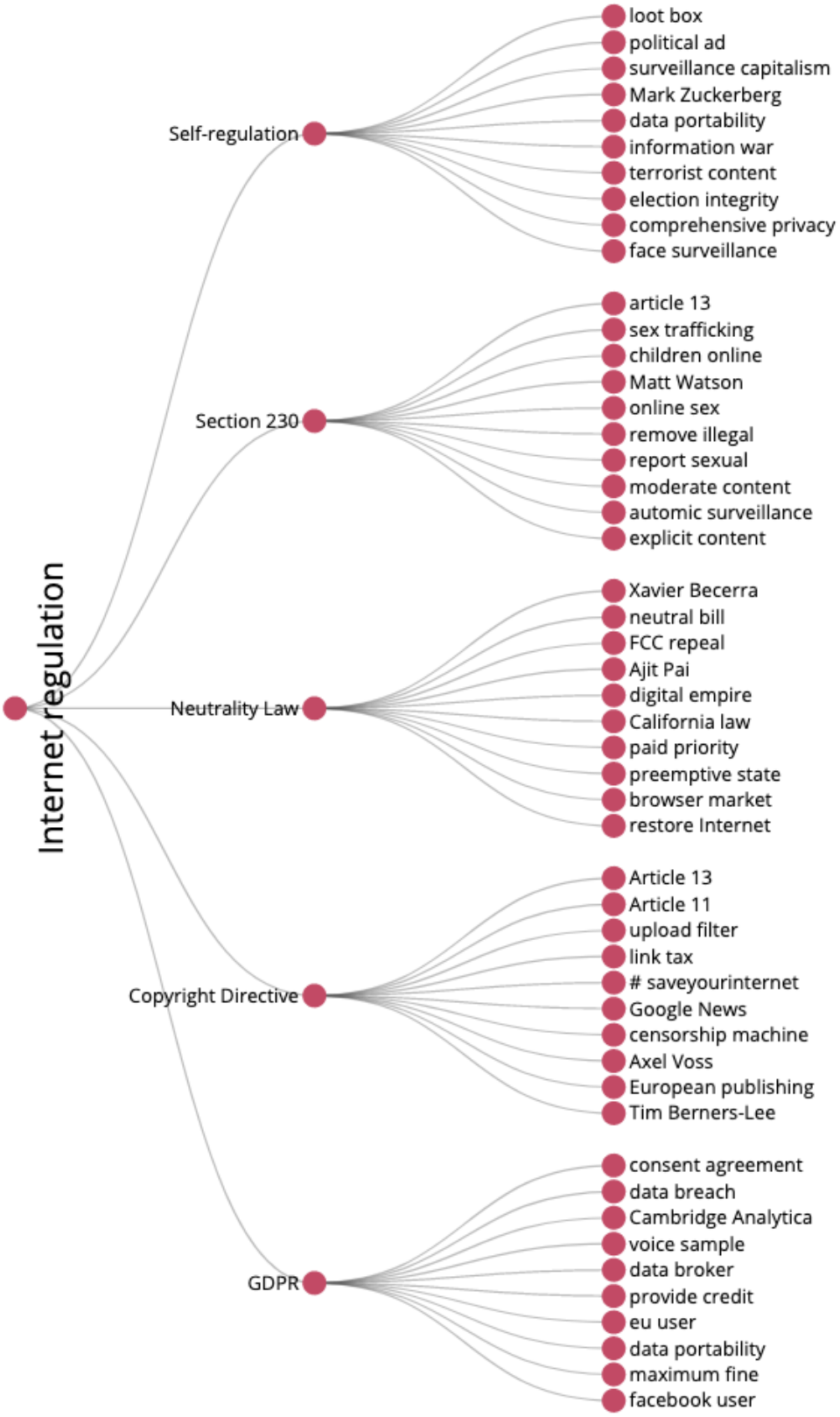


Figure 26. Sentiments analysis: GDPR

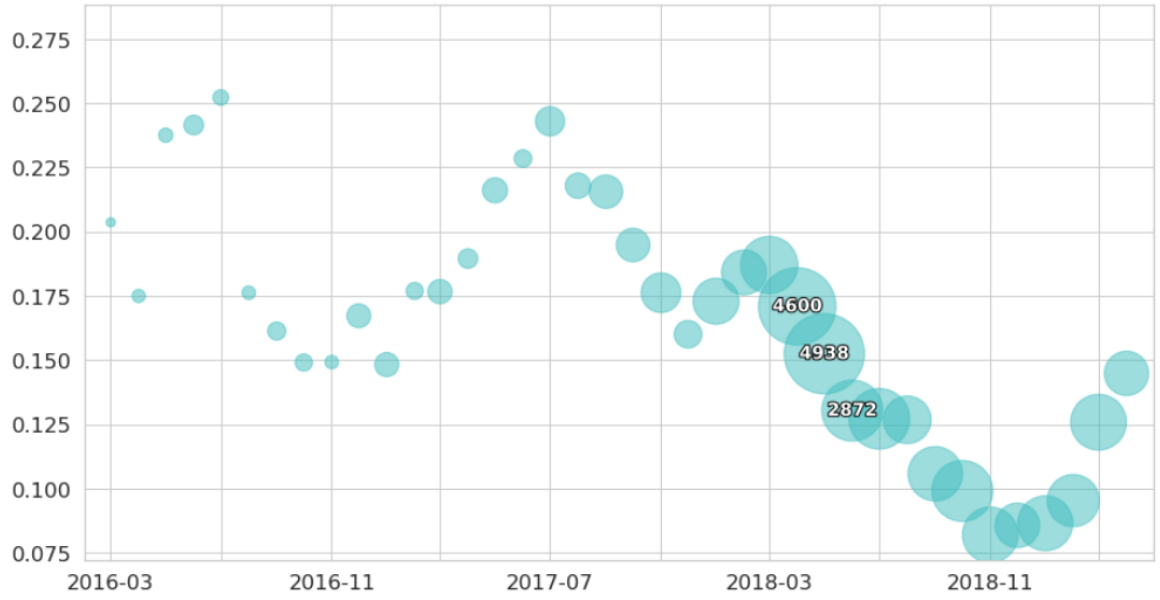


Table 13. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
voice assistant ai research digital marketing third-party data processing data	british airways article 11 electronic health facebook data plain text

Figure 27. Sentiments analysis: copyright directive

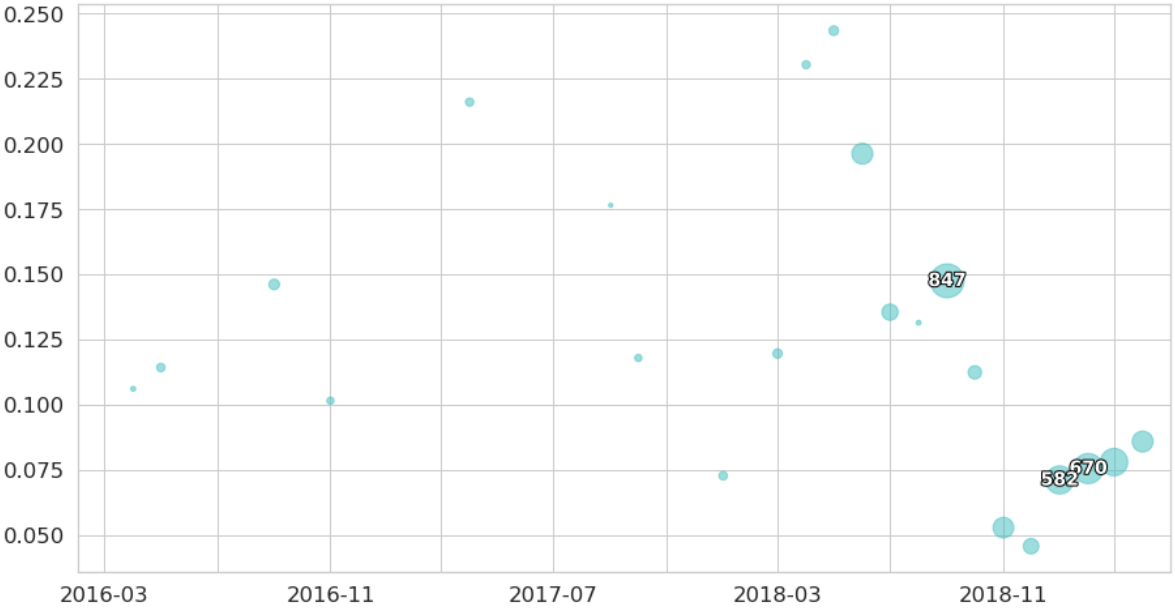


Table 14. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
european creators fair remuneration creative content #saveyourinternet european publishers	automated surveillance committee voted recognition technology susan wojcicki key legislative

Table 15. Topic modeling: Internet regulation

Topic 2 GDPR (13.3% of tokens)	Topic 3 Smart city (7% of tokens)	Topic 4 Ridesharing apps (4.3% of tokens)
court law government information facebook protection gdpr legal rights enforcement	car smart robot city wi-fi health dns home vehicles sensors	uber trump tesla drivers copyrights google china self-driving lawsuit tax
Topic 6 Net neutrality (2% of tokens)	**Topic 7 Copyright reform (1.3% of tokens)**	**Topic 14 Surveillance (0.6% of tokens)**
fcc neutrality net pai rules internet isps broadband vote repeal	european copyright commission parliament directive content publishers digital platforms rules	nsa 702 wikileaks assange surveillance snowden banking intelligence cia fisa

5.8 Chinese tech sector

Observing China's efforts towards becoming a major internet superpower has clearly taught the Western countries that technology is not neutral. On the contrary, allegations against Chinese tech giant Huawei regarding spying, theft of intellectual property, obstruction of justice and fraud, as well as controversies concerning Google's Dragonfly project revealed to what extent Internet is political.

The co-occurrences for Chinese tech show that the discussion was focused around major tech giants, such as Huawei (world's #1 telecom supplier and #2 phone manufacturer) or ZTE,

prominent figures of Chinese internet sector as Huawei's CEO Ren Zhengfei, CFO Meng Wanzhou and former Google China president Kai-Fu Lee. The former two denied the allegations of the U.S. government that the company provides a backdoor to Chinese intelligence (see: state-sponsored espionage co-occurrence). Wanzhou was arrested in Canada, where she is facing Iran fraud charges[57]. Kai-Fu Lee famously argued in his recent book that China is moving forward to become the global AI leader and may surpass the United States, exploiting the big data of its population[58].

In the light of cyber espionage allegations, the US Commerce Department's Bureau of Industry and Security effectively banned Huawei and all of its subsidiaries from US communications networks. Five Eyes - an anglophone intelligence alliance comprising of USA, Australia, Canada, New Zealand and the UK has announced that they will not use technology from Huawei in the "sensitive" parts of their 5G telecom networks (see: five eyes co-occurrence)[59].

Google's controversial Dragonfly project has appeared frequently in the discussions about Chinese government. The project is a prototype search engine compatible with Chinese censorship provisions. After the employees protests, the Dragonfly project was reportedly shut down. However, according to some sources, work on it still carried on in 2019[60].

Articles covering the Chinese tech sector have grown in number since 2018. This increase of coverage has been paired with a decline in sentiment, suggesting an increase of negative news stories. The co-occurring words that were in the paragraphs with the most negative sentiment scores reflect the recent scandals involving Huawei, e.g. the arrest CFO Wanzhou. "Uncle Sam" is referring to the US government that has been actively blocking the US expansion of Huawei[61].

On the other hand, the coverage of Chinese tech sector has been rather positive in the context of AI development, which is one of China's technological advantages[62].

The topic modeling confirmed that media coverage of Chinese tech industry was particularly focused on the so called tech trade war, resulting in banning Huawei from US communications networks. Other important topics were i.a. Chinese efforts to curb greenhouse gas emissions and developments in automation technologies.

Conclusions: Not long time ago the Internet was considered to be a force for greater democratization. However, the example of China shows how authoritarian regimes can use the Internet to control societies. The expansion of Chinese tech sector in the Western world raises crucial cybersecurity questions.

[57] https://arstechnica.com/tech-policy/2019/01/us-asks-canada-to-turn-over-huaweis-cfo-on-alleged-sanctions-violations/
[58] Kai-Fu-Lee, *AI Superpowers: China, Silicon Valley, and the New World Order*, 2018.
[59] https://www.theguardian.com/technology/2019/apr/30/alleged-huawei-router-backdoor-is-standard-networking-tool-says-firm
[60] https://theintercept.com/2019/03/04/google-ongoing-project-dragonfly/
[61] https://www.theregister.co.uk/2018/12/06/huawei_cfos_arrest/
[62] https://www.theverge.com/2019/3/14/18265230/china-is-about-to-overtake-america-in-ai-research

Figure 28. Co-occurrence analysis for Chinese tech

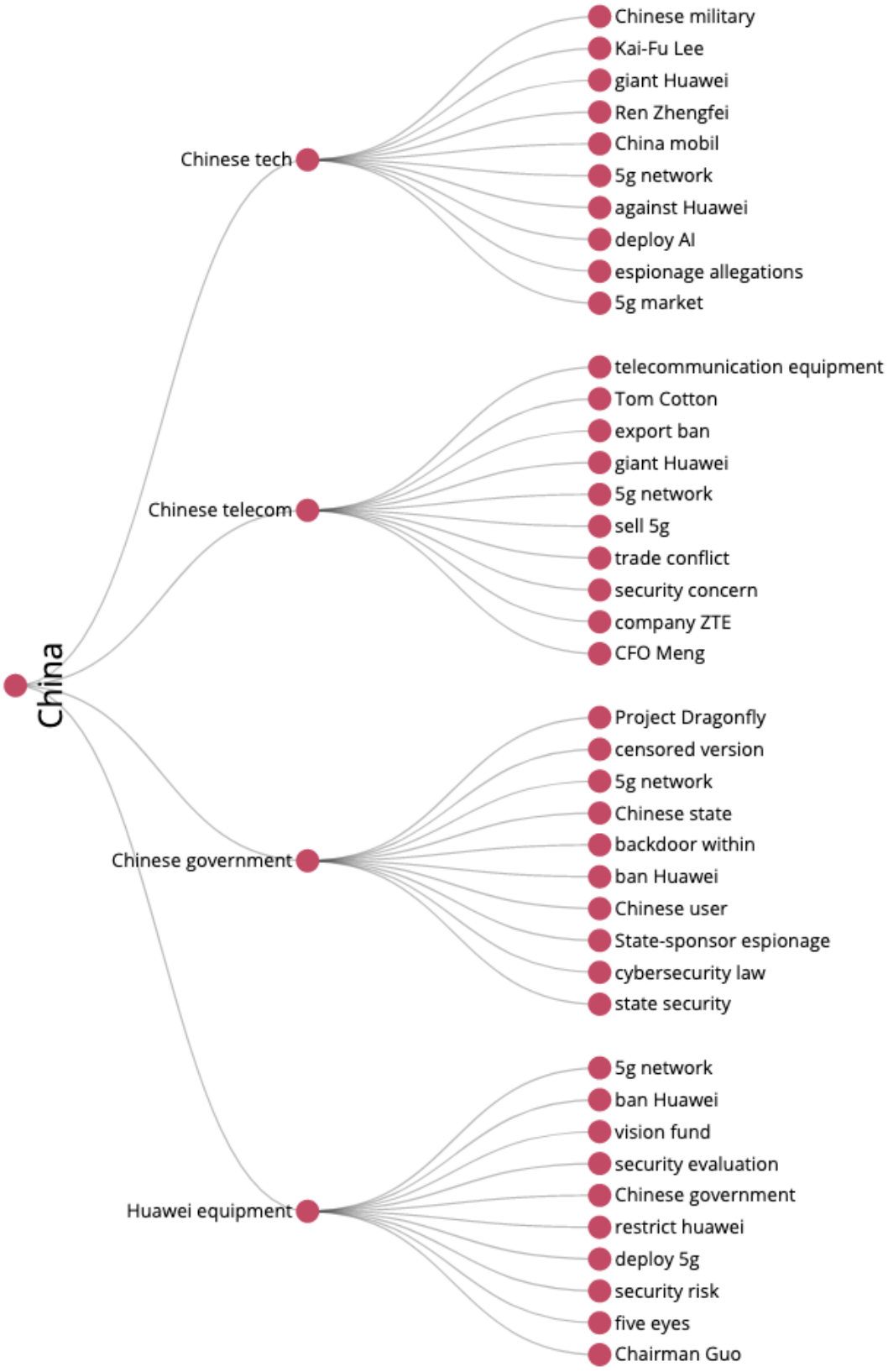


Figure 29. Sentiments analysis: Chinese tech

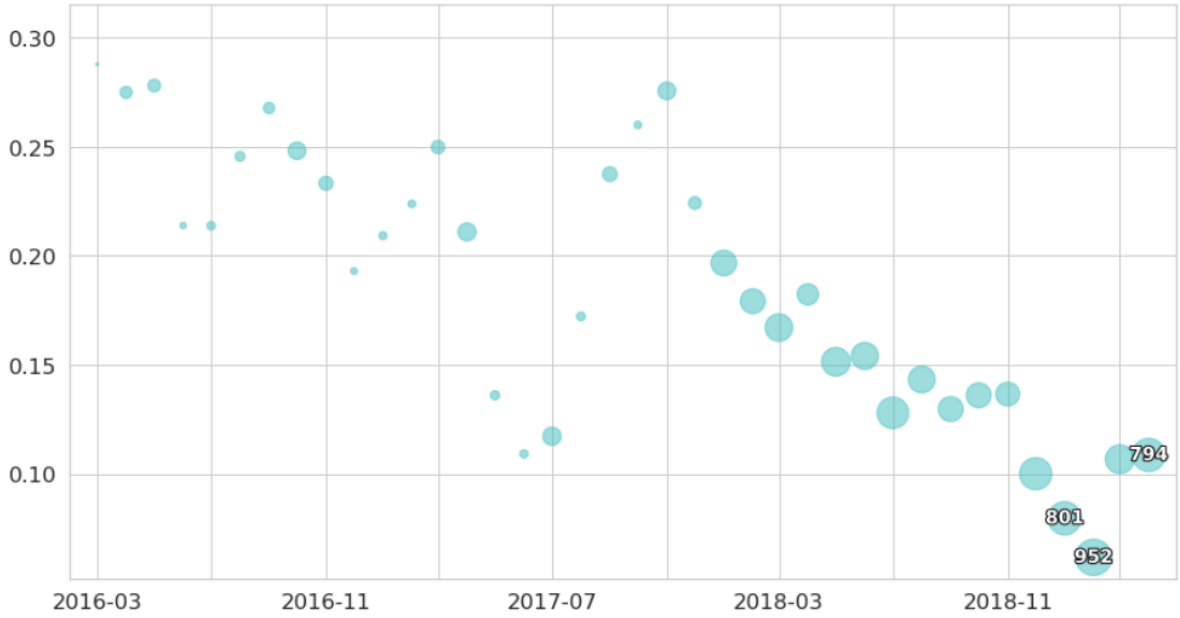


Table 16. Co-occurrences with most positive and negative sentiments

Most positive	Most negative
ai platform ai technology ai applications ai algorithms neural network	uncle sam meng wanzhou founder ren huawei employee huawei products

Table 17. Topic modeling: Chinese tech

Topic 2 Huawei ban (30% of tokens)	Topic 3 Transportation (7.7% of tokens)	Topic 6 Energy (1.6% of tokens)
government	uber	energy
security	alibaba	climate
huawei	didi	quantum
trump	tencent	solar
u.s.	investment	carbon
data	capital	emissions
companies	e-commerce	renewable
national	softbank	coal
internet	startup	water
report	funding	wind
Topic 8 Crypto (1.5% of tokens)	**Topic 19 Automation (0.5% of tokens)**	**Topic 21 Apps (0.5% of tokens)**
bitcoin	robots	tiktot
zte	automation	bytedance
blockchain	boeing	musical.ly
cryptocurrency	aircraft	toutiao
alphago	ctrip	douyin
icos	ubtech	e-sports
trading	autonomous	duanzi
bank	humanoid	gps
mining	colocation	videos
ban	irobot	aggregator

6 Conclusions

This analysis applied a methodology for identifying trending topics in online tech media and academic working papers enabling a deeper exploration of internet technologies and related social challenges.

We have proposed a sequential text mining framework well suited for informing about the fast changing tech landscape. Our methodology brings together a set of straightforward text mining exercises that are easy to diagnose, tune, evaluate and interpret. The sequence of methods enables the exploration of technology related texts by different levels of granularity. The terms frequency analysis provides a birds eye view on the emerging technologies and interrelated social issues. The co-occurrence analysis helps building the topologies of the most relevant topics. The changing public perception is tracked by the sentiment analysis. The combination of the co-occurrence and sentiment analysis is used to unravel the positive and negative stories related to a topic. Finally, topic modeling provides a robustness check to identify dominant themes of discussion.

The potential of the methodology has been demonstrated with deep dives on eight case studies. In order to present the possibilities of the method, the selected topics concern frequently discussed technologies and regulatory challenges. Therefore, the report provides insights into widely defined topics, mapping fields, key actors and other related issues.

However, these tools can be implemented for the investigation of narrow topics as well, including technological subfields and clearly defined social problems.

While this report provides a description of results, the presented visualisations are static versions of the presentation available online (https://fwd.delabapps.eu/). The results can be explored in interactive applications, enabling a deeper analysis. Moreover, the raw results are publicly available online at the Zenodo platform (https://zenodo.org/communities/ngi_forward), facilitating further research.

Appendix

Top 20 most trending unigrams and bigrams

Following the steps discussed in the methodology section, the lists of unigrams and bigrams are sorted by the regression coefficient and NGI-related terms are manually selected.

Table A1. Top trending bigrams

#	Word
1	cambridge analytica
2	fake news
3	facial recognition
4	5g network
5	tech giant
6	discuss subject
7	iphone x
8	climate change
9	iphone xs
10	industrial leader
11	hear industry
12	leader discussion
13	mark zuckerberg
14	expo world
15	world series
16	big tech
17	cyber security
18	google assistant
19	pixel 3
20	neural network

Table A2. Top trending unigrams

#	Word
1	facebook
2	2018
3	ai
4	5g
5	huawei
6	2019
7	china
8	chinese
9	elect
10	fake
11	cambridge
12	analytica
13	ban
14	zuckerberg
15	climate
16	instagram
17	cryptocurrency
18	gdpr
19	expo
20	musk

Source weights

The weights used in the frequency analysis have been assigned based on the number of articles published by the website and whether a source is of European origin.

To illustrate: the weight of The Conversation is 5% considering the low number of articles, while The Guardian and Register receive above average weights of 12% as a counterweight to reduce the dominance of US based sources. The weights are the following:

Table A3. Source weights

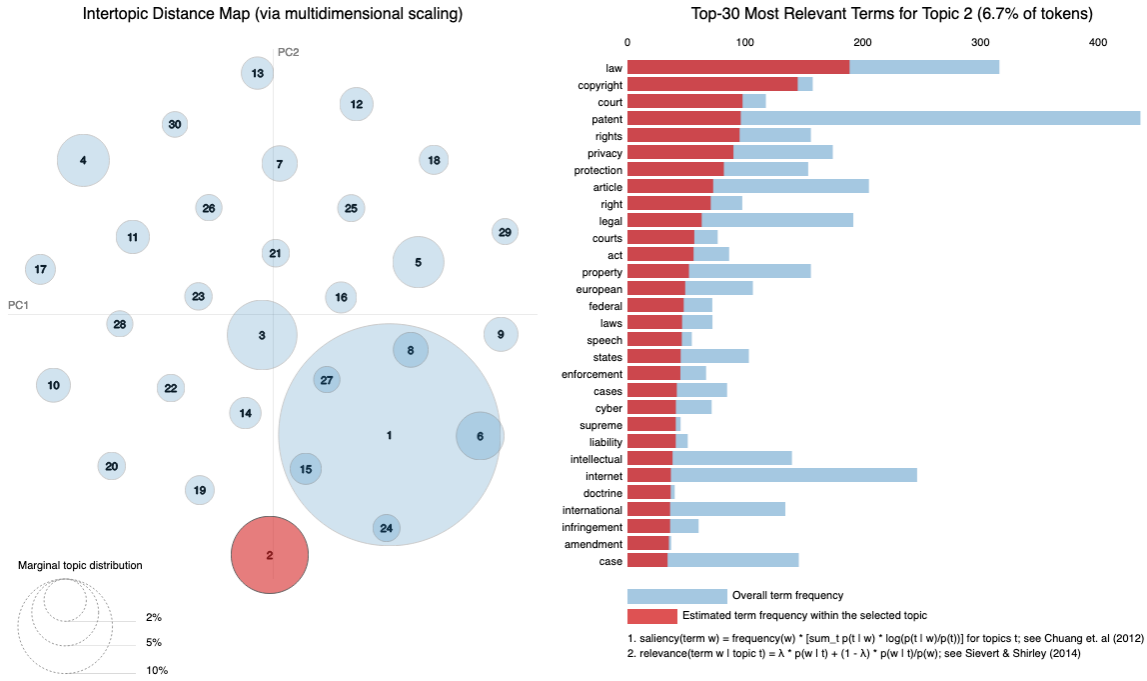
Source	**Weight**
Arstechnica	0.05
Euractiv	0.05
Fastcompany	0.05
Gizmodo	0.09
IEEE Spectrum	0.05
Politico Europe	0.05
Reuters	0.05
Techcrunch	0.09
Techforge	0.05
The Conversation	0.05
The Guardian	0.12
The Register	0.12
The Verge	0.09
ZDNet	0.09

Table A4. Keywords used to filter relevant articles in the topic modeling

Umbrella topic	**Keywords**
AI and ML	'ai', 'facial', 'maven', 'reinforcement', 'neural', 'machine-learning', 'algorithm', 'ml', 'robot', 'black-box', 'pytorch', 'ai-driven', 'duplex'
IoT	'iot', 'iiot', 'bgp', 'ar/vr', 'co-loc'
Quantum computing	'quantum' , 'qubit', 'd-wave'
Blockchain and cryptocurrencies	'blockchain','cryptocurrency', 'ico', 'tether', 'coinbase', 'monero', 'bitfinex', 'stablecoin', 'decentralisation', 'blockchain-based', 'cryptoassets'
Internet regulation	'230', 'neutrality', 'self-regulation', 'copyright', 'fcc', 'loot', 'privacy', 'gdpr'
Social media and content crisis	'conspiracy', 'content', 'troll', 'disinformation', 'censorship', 'fact-check', 'deepfake', 'bot', 'infowars', 'lakhta', 'dragonfly'

Market competition	'market', 'competition', 'gafa', 'monopolist', 'gig', 'antitrust'
Chinese tech sector	'weibo', 'chinese', 'china', 'huawei'

Figure A1. Topic modeling - SSRN



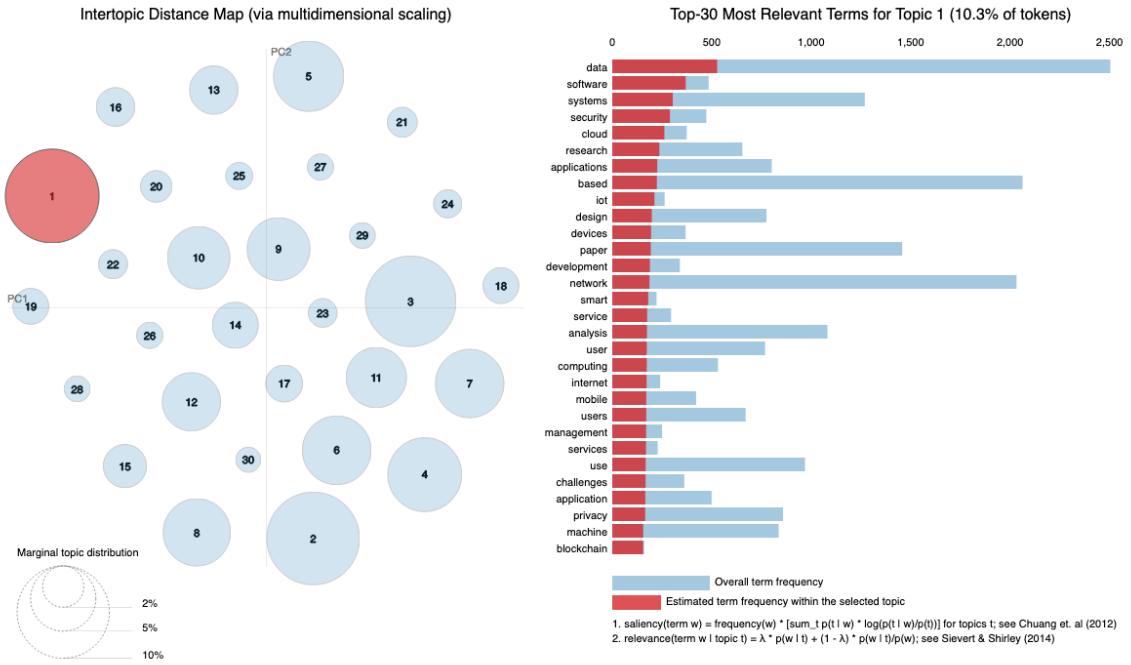
Note: Interactive visualization tool available at: https://fwd.delabapps.eu/

Table A5. Topic modeling - SSRN: Top-10 keywords in selected topics

Topic 2 Regulations (6.7% of tokens)	Topic 6 IoT (2.6% of tokens)	Topic 9 Content crisis (1.3% of tokens)
law copyright court patent rights privacy protection article right legal	iot network web semantic cloud sensor wireless algorithm ontology computing	media news sentiment twitter fake social icann internet governance facebook
Topic 10 Patents (1.3% of tokens)	**Topic 11 Cloud (1.3% of tokens)**	**Topic 14 Blockchain & crypto (1.1% of tokens)**
patent licensing frand standard royalty	cloud computing algorithmic server contracts	blockchain bitcoin cryptocurrencies fintech money

litigation sep damages infringement antitrust	collusion confidentiality ideal smart arbitrage	financial payment distributed ledger technology

Figure A2. Topic modeling - ArXiv



Note: Interactive visualization tool available at: https://fwd.delabapps.eu/

Table A6. Topic modeling - ArXiv: Top-10 keywords in selected topics

Topic 4 5G (6.4% of tokens)	Topic 7 Recommendation systems (5.5% of tokens)	Topic 10 Reinforcement learning (4.6% of tokens)
channel	social	learning
wireless	user	reinforcement
power	data	policy
energy	research	deep
mimo	recommendation	agent
interference	media	training
communication	information	model
network	online	task
transmission	content	neural
rate	news	search
Topic 16 Robotics (1.8% of tokens)	**Topic 18 Cybersecurity (1.5% of tokens)**	**Topic 25 Privacy (0.9% of tokens)**
robot	adversarial	privacy
motion	attacks	differential
camera	malware	fingerprint
navigation	detection	preserving
localization	examples	signature
robots	robustness	encryption
depth	perturbations	password
sensor	anomaly	homomorphic
estimation	defense	utility
human	deep	leakage