D1.12: Final open data sets & code, and methodology guides

Work package	WP1
Task	Task 1.1 Collection of online data & data-set management Task 1.2 Continuous identification of new technologies and trends Task 1.3 Topic co-occurrence analysis: where do technology and social issues meet? Task 1.4 Network Analysis through Topic modeling & deep learning algorithms
Due date	31/12/2021
Submission date	30/12/2021
Deliverable lead	UNIWARSAW
Dissemination level	Public
Nature	Report / Open Research Data Pilot
Authors	Kristóf Gyódi Michał Paliński Łukasz Nawaro Katarzyna Śledziewska Maciej Wilamowski

Version	**1**
Reviewers	**Alberto Cottica, Ida Nissen**
Status	**Final**

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 N°825652**

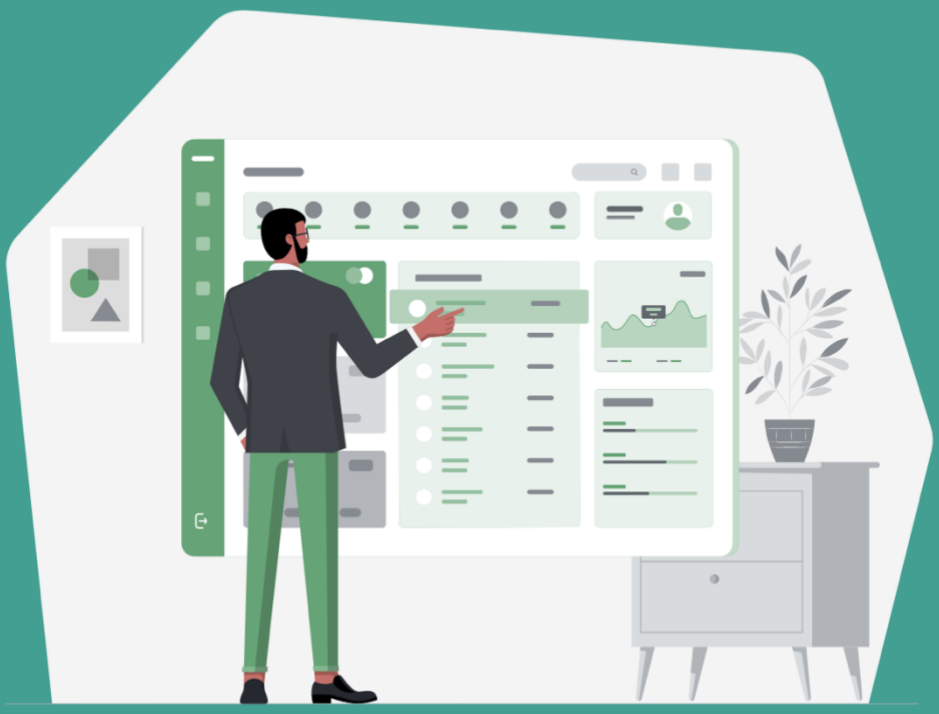
Abstract

This report provides an overview of the resources and output produced by the DELab team at the University of Warsaw during the Next Generation Internet (NGI) Forward project. The aim of this document is to describe and collect all pieces of work in one place, facilitating their continued use following the project. As part of NGI Forward, our main aim has been to support the Next Generation Internet initiative by providing data science tools to map and analyse the developments of the tech world. Therefore, there are different types of outputs: methodologies with documented Python codes; reports summarizing insights; and interactive presentations with publicly available data. We briefly introduce and explain the two main text-mining methodologies (trend analysis and topic mapping) and novel datasets prepared during the project, as well as present the list of available reports, tutorials and other project results. The project materials are also presented in a website: https://fwdmain.delabapps.eu.

NGI FORWARD

NGI FORWARD: OVERVIEW OF DATA, METHODS AND CODES

December 2021



NEXT GENERATION INTERNET
INTERNET OF HUMANS

AUTHORS

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, Katarzyna Śledziewska, Maciej Wilamowski

---

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 N°825652

Introduction

This report provides an overview of the resources and output produced by the DELab team at the University of Warsaw during the Next Generation Internet (NGI) Forward project. The aim of this document is to describe and collect all pieces of work in one place, facilitating their continued use following the project.

As part of NGI Forward, our main aim has been to support the Next Generation Internet initiative by providing data science tools to map and analyse the developments of the tech world. More precisely, we have focused on three goals:

- To develop text-mining methodologies to extract insights on issues relevant to NGI
- To prepare case studies highlighting key conclusions from the data-driven research
- To publish the results in forms facilitating further use and research

Therefore, there are different types of outputs: methodologies with documented Python codes; reports summarizing insights; and interactive presentations with publicly available data.

First, we briefly introduce and explain the different text-mining techniques. We have prepared two complementary methodologies: *trend analysis* and *topic mapping*. Trend analysis serves to identify emerging terms in news articles: the methodology captures among others technologies, social and regulatory issues, events and persons gaining relevance over time. Moreover, the presented pipeline contains further steps for the analysis of trending issues, such as the exploration of sentiments. On the other hand, topic mapping is a more suitable methodology in case the user would like to gain insights on specific areas of interest, such as privacy, ethical AI or climate change.

The two different goals and methodologies required different input data. For the analysis of trends, complete datasets of documents are necessary, such as all articles published by a website in a given period of time. We compiled a dataset of 14 major English-language technology websites from the US, EU and Australia for the period of 01.2016 - 04.2021 (the *online news* dataset). In the case of the topic mapping exercise, however, there is no need to track all articles published at a given time, instead the emphasis is on the diversity and relevance of articles. For this reason, we collected documents shared on different social media platforms and different languages (the *social media* dataset).

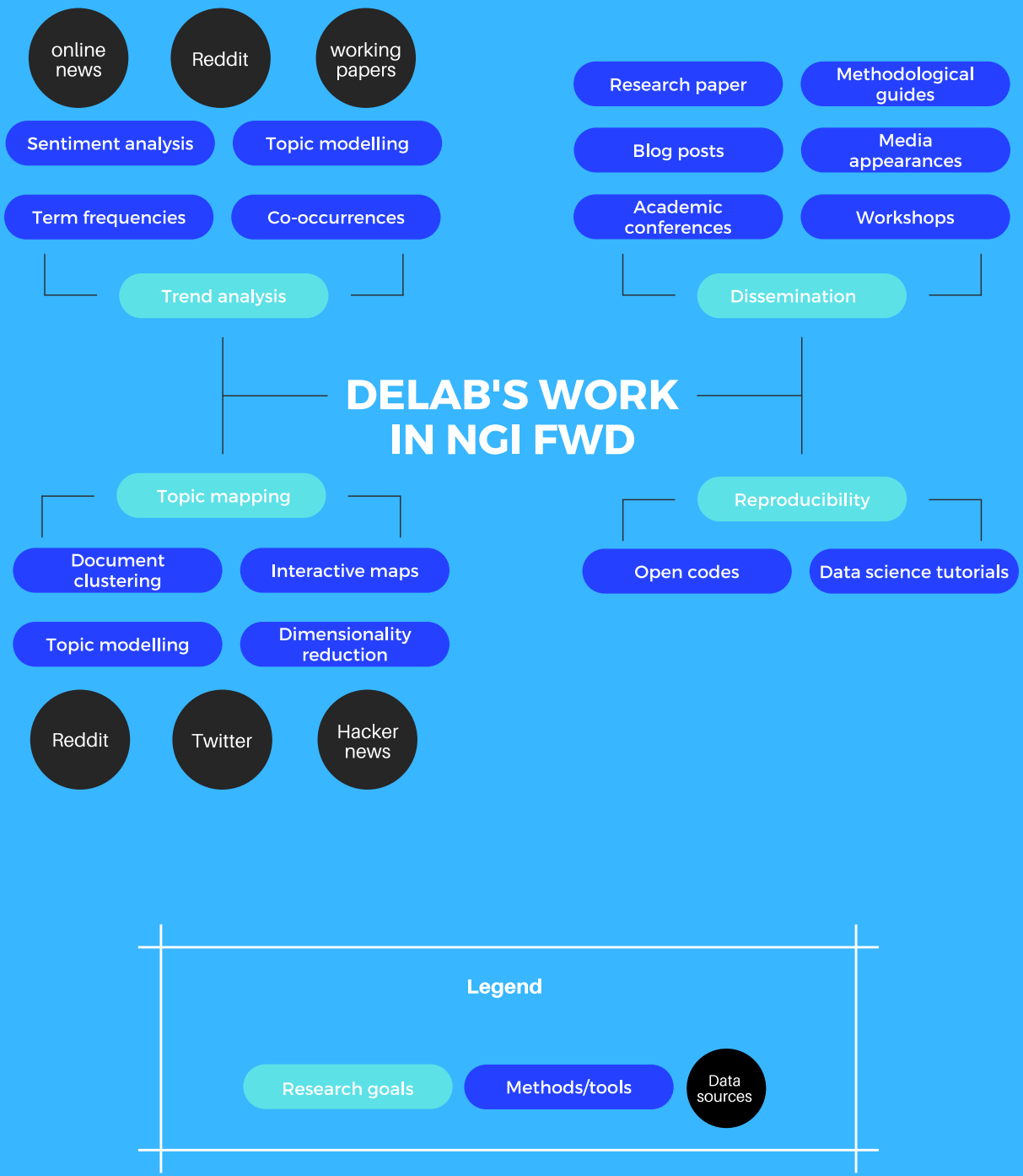
Based on these two methodologies, we prepared various reports and presentations. In the first phase of the project, our efforts centered on the identification of trending issues in tech media articles: two presentations were prepared on major emerging technology and policy areas, and an analysis focused

on the COVID-19 pandemic. We also published a preliminary topic mapping work that identified the main topics discussed in the online news dataset.
In the second half of the project, the topic mapping analysis was dominant. Two major works were prepared: the first explored English-language articles shared on social media, while the second complemented the results with the analysis of discussions in German, Polish, Spanish and Portuguese, hence including perspectives both from Europe, as well as from Latin America.

We also put emphasis on the reproducibility of the results. The codes to collect the data and prepare the trend analysis and topic mapping are published on Gitlab, along with a tutorial on the use of the codes.

Finally, this guide ends with a list of project outputs with relevant links and descriptions.
The project materials are also summarised online in a main website: https://fwdmain.delabapps.eu.

Figure 1. The scheme of our output



Methods

Throughout Next Generation Internet Forward, we have been pursuing two complementary approaches to examine technology and policy issues with the help of text-mining techniques. Our first proposition focused on the identification of emerging issues gaining importance over time: we refer to this methodology as *trend analysis*. The aim of the second approach is to map specific topics discussed in online documents - we call it *topic mapping*. The two approaches have a key difference: in trend analysis we do not have any prior assumptions on what topics we want to analyse, while for topic mapping a preliminary selection of broad areas of investigation are necessary. As an example, if we are interested in gaining insights on a defined problem, such as privacy or climate change, topic mapping provides more suited solutions. On the other hand, if we would like to highlight issues gaining traction, trend analysis provides the necessary tools. Let us briefly explain the main steps and assumptions of the two methodologies.

Trend analysis

The main aim of trend analysis is to capture terms gaining traction in text documents. Additionally, our methodology also includes further tools to map and explore emerging topics. The trend analysis consists of four main steps:

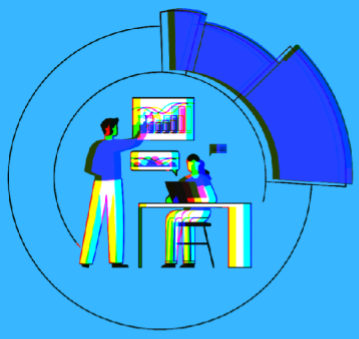
- Term frequencies
- Co-occurrences
- Sentiment analysis
- Topic modelling

The proposed pipeline enables the exploration of issues at different levels. First, we identify trending terms based on the analysis of term frequencies. The most trending terms can be highlighted using regression analysis. These identified terms are the basis for further analysis. The co-occurrence analysis presents pairs of terms that are most frequently mentioned together, finding connections between relevant topics. In order to track the public perception of issues and identify the positive and negative news stories related to a selected topic, sentiment analysis is performed. To additionally verify our results, topic modelling is prepared using Latent Dirichlet Allocation (LDA). LDA shows which terms define key topics across the text corpus, providing information on broad topics and co-occurrences. In practice, LDA helps understanding the hidden (latent) topics in discussions.

In the next sections we briefly introduce the various steps, with a focus on the methodology assumptions. To illustrate the process, we use the results from the trend analysis in 2020[1].

[1] https://fwd2020.delabapps.eu

Figure 2. Main steps of trend analysis



TERM FREQUENCY

Identify changes in frequencies over time

Regression analysis and expert selection

CO-OCCURRENCES

Find terms co-occurring in the same texts

Co-occurrence matrices and expert selection

SENTIMENT ANALYSIS

Quantify the polarity of a given term

Lexicon & rule-based sentiment analysis and expert selection

TOPIC MODELLING

Reveal and map latent topics in a corpus of texts

Latent Dirichlet Allocation

Term frequencies

In trend analysis we assume that there is a significant relationship between the importance of an issue and the frequency it is being mentioned. As an example, if the term "5G" is relatively more frequently mentioned (e.g. its share among all terms increased) over time, it indicates that news articles cover more intensively issues related to 5G. Therefore, the basis of trend analysis is the calculation of term frequencies: for each term and source, the average monthly term frequency has been calculated by dividing the number of occurrences of the term by the number of occurrences of all terms. Besides the frequency of single terms (unigrams), we also calculated the frequency of bigrams - two terms next to each other in the text (e.g. 'fake news'), discovered automatically by *gensim*'s Phrases[2].

An important assumption we make is that all terms are equally important in the text. Alternatively, we can also consider the position of the term in the text. As a robustness check, we have calculated the results with more weight assigned to terms located in titles and in the first paragraphs, however, the results did not support significant advantage[3].

Afterwards, the weighted average of frequencies by source has been calculated. Therefore, we make further assumptions related to weighting the different sources in an interpretive step of the analysis. There are significant differences in the number of articles published across the sources: the smallest publisher accounts for less than 1% (*Euractiv*) of all articles, while the largest makes up more than 20% (*TechCrunch*). Does this mean that a source is 20 times more relevant than another? Similarly, the majority of articles in our dataset are published by websites in the US, while the smaller publishers of the dataset are based in the EU. On the other hand, the debate in European media may be more relevant for the project. For these considerations, we have assigned weights to ensure that smaller publishers have a greater influence on the results than their share of publications would determine[4]. We also tested results for the alternative method of assigning equal weight for all sources and found that weighted sources perform better than unweighted[5].

Following the calculation of the weighted term frequencies, we identified the most trending terms with a regression analysis: 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 for each term that highlights the direction

[2] https://radimrehurek.com/gensim/models/phrases.html
[3] https://policy.delabapps.eu/#OB4
[4] https://zenodo.org/record/5788431
[5] https://policy.delabapps.eu/#OB4

(increase or decrease) and magnitude in the changes in frequencies over time.

Next, the terms were sorted based on the coefficient, enabling the identification of the most trending terms. However, the top growing words are often stopwords due the large number of occurrences. To exclude irrelevant terms in an automatic way, a normalised coefficient is used (coefficient divided by the mean weighted frequency in all months of the regression) with a threshold value required for terms to achieve.

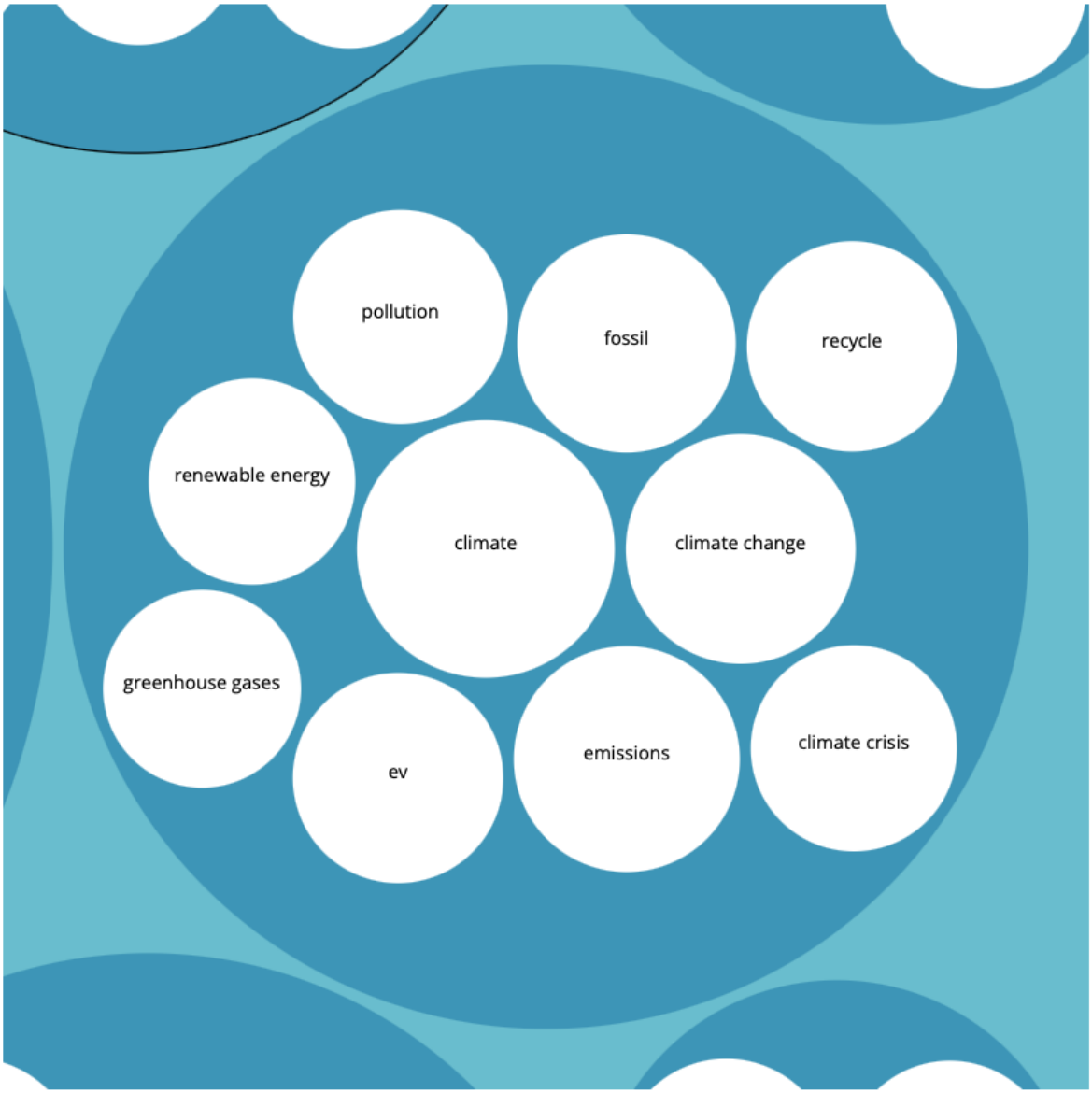
Therefore, the first algorithmic stage of the analysis ends with identifying the most significantly growing terms (with the largest coefficient which are above the threshold for normalised coefficient).

The next step is interpretive: we receive a list of trending terms, highlighting technologies, companies, events, products, and policy proposals. During each iteration of preparing the analysis (we have updated the results multiple times during the project), we reviewed the top 1000 most trending terms, selected the ones most relevant to Next Generation Internet, and grouped them according to themes. Based on the discovered areas, we identified a number of key umbrella topics in each iteration of the analysis. This is a subjective part of the study: other researchers may interpret the results differently. However, the raw results have been published at Zenodo[6], enabling the analysis outside our team.

Figure 3. Identified key umbrella topics in 2020

[6] https://zenodo.org/record/5788431

Figure 4. Example for trending terms related to climate change



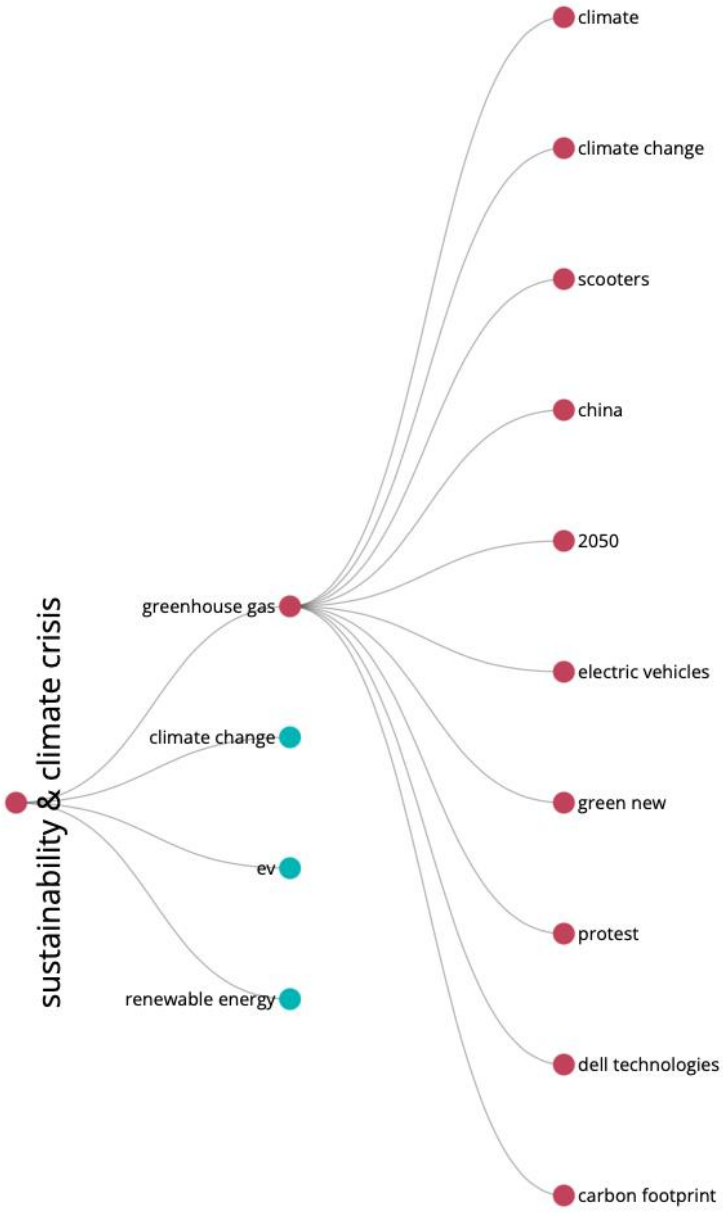
Co-occurrence analysis

Co-occurrences were calculated for selected terms from the previous list of trending terms. The selection of terms was an interpretive process: those terms were selected that are representative of the general umbrella topics. For these terms (“analysed terms”), the most common “co-occurring" terms out of the top 15000 most significantly growing terms have been calculated. The terms co-occur if they are located in the same article, it is not required to be in the same sentence or paragraph. First, the number of occurrences of the co-occurring term in all articles in a given source containing the analysed term has been counted. This number of co-occurrences was divided by the

number of occurrences of the analysed term in all articles in the source. Sources have been aggregated using the same weights as in the case of term frequencies.

Similarly to the analysis of term frequencies, the process ends with lists of co-occurring terms, sorted by the co-occurrence scores. Following this algorithmic process, interpretive analysis decides which terms among the top co-occurrences are summarised in the presentations. Therefore, the final results reflect our subjective opinions. However, we also published all raw results at Zenodo for further analysis[7].

Figure 5. Example for co-occurrences



[7] https://zenodo.org/record/5788439

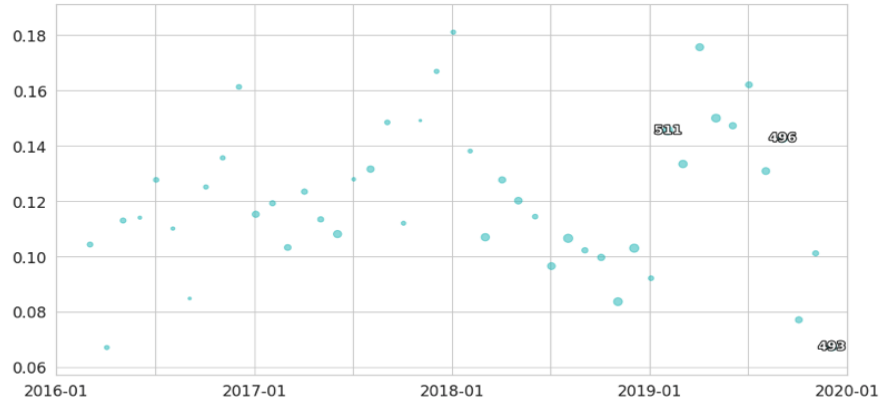
Sentiment analysis

Sentiment analysis is a field of text-mining focused on identifying the emotions in documents. We are interested in extracting the general attitude towards policies and technologies: our aim is to determine whether a subject is discussed in a rather positive or negative way. The sentiment analysis has been prepared using VADER[8] , an open-source lexicon and rule-based sentiment analysis tool. As VADER is more robust in the case of shorter social media texts, the analysed articles have been divided into paragraphs. VADER assigns a score for paragraphs with values between -1 (extremely negative) and +1 (extremely positive).

The same words which have been chosen for the co-occurrence analysis were selected for the sentiment analysis. All paragraphs in the articles containing the given term were modified to exclude the term - it was necessary as the term itself may not be emotionally neutral.
We presented two analyses: the average monthly sentiment of paragraphs containing the selected terms, 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 were selected. The sentiment was computed on paragraphs modified by removing both the analysed and frequently co-occurring terms. Based on the average sentiment scores of paragraphs, the co-occurring terms with the most negative and positive sentiment were identified. Similarly to the previous steps, the final most positive and most negative co-occurrences were selected in an interpretive process.

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

Figure 6. Example for sentiment analysis for the term “greenhouse gas ”(up - average sentiment scores over time, down - most positive / negative co-occurrences)



GREENHOUSE GAS

Most positive	Most negative
heat pumps	geoengineering
autonomy	climate emergency
plantbased	extinction
solar wind	protest
electric vehicles	democracy

Topic modelling

The first three steps of the analysis provide a funnel: from all terms present in the documents, the trending ones are captured and the relationships between them are explored. Topic modelling is orthogonal to these steps: it can be omitted from trend analysis. The aim of topic modelling is the general identification of topics in documents, irrespective whether the topic is trending over time. Therefore, topic modelling is more suited to the second set of methods presented by us as topic mapping. In fact, the experimentation with topic modelling provided the foundations for the creation of the topic mapping pipeline. We have included topic modelling in the trend analysis as an additional step that may capture insights not picked up by the previous tools.

The following sections are cited from our topic modelling report from 2019[9].
Topic modelling assumes that documents, such as news articles, consist of words which can be assigned to various distinguishable topics. As an example, a news article covering the Cambridge Analytica scandal may contain the following topics: social media, politics and tech regulations, with the following proportions: 60% social media, 30% politics and 10% tech regulations. The other assumption is that topics contain characteristic vocabularies,

[9] https://fwd.delabapps.eu/topic_modelling.html

e.g. the social media topic is described by the words Facebook, Twitter etc.
Latent Dirichlet Allocation (LDA) is a popular topic modelling method due to its ease of use, flexibility and interpretable results. LDA has been proposed by Blei et al. (2003)[10], based on Bayesian statistics. The method's name provides its key foundations. Latent comes from the assumption that documents contain latent topics that we do not know a priori. Allocation shows that we allocate words to topics, and topics to documents. Dirichlet is a multinomial likelihood distribution: it provides the joint distribution of any number of outcomes. As an example, Dirichlet distribution can describe the occurrences of observed species in a safari (Downey, 2013)[11]. In LDA, it describes the distribution of topics in documents, and the distribution of words in topics.
Documents in the corpus are treated as bag-of-words, i.e. the word ordering is not taken into account. The premise of the topic modelling methods is simple: we describe and recreate our texts with a combination of topics consisting of specific words. More precisely, we aim at recreating our word-document matrix with the combination of two matrices: the matrix containing the Dirichlet distribution of topics in documents (topic-document matrix), and the matrix containing the words in topics (word-topic matrix). The construction of the final matrices is achieved by a process called Gibbs sampling. The idea behind Gibbs sampling is to introduce changes into the two matrices word-by-word: change the topic allocation of a selected word in a document, and evaluate if this change improves the decomposition of our document. Repeating the steps of the Gibbs sampling in all documents provides the final matrices that provide the best description of the sample.

Three parameters need to be provided by the user: the number of topics the algorithm is supposed to find; the prior topic-document distribution and the prior word-topic distribution.
To identify the appropriate parameters, we used coherence measures proposed by Röder (2015)[12]. The method assesses the coherence of topics by analysing the pairs of the most frequent topic-words. The coherence approach favours topics with words semantically similar to each other, facilitating the interpretability of the results. An intuitive explanation for coherence is provided in a blog post by Kapadia (2019)[13].

[10] Blei, D., M., Edu, B., B., Ng, A. Y., Edu, A., S., Jordan, M., I., Edu, J., B. (2003), Latent Dirichlet Allocation. Journal of Machine Learning Research 3.

[11] Downey, A., B., (2013), "Think Bayes", O'Reilly Media Inc, USA.

[12] M. Röder, A. Both, A. Hinneburg. (2015), 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.

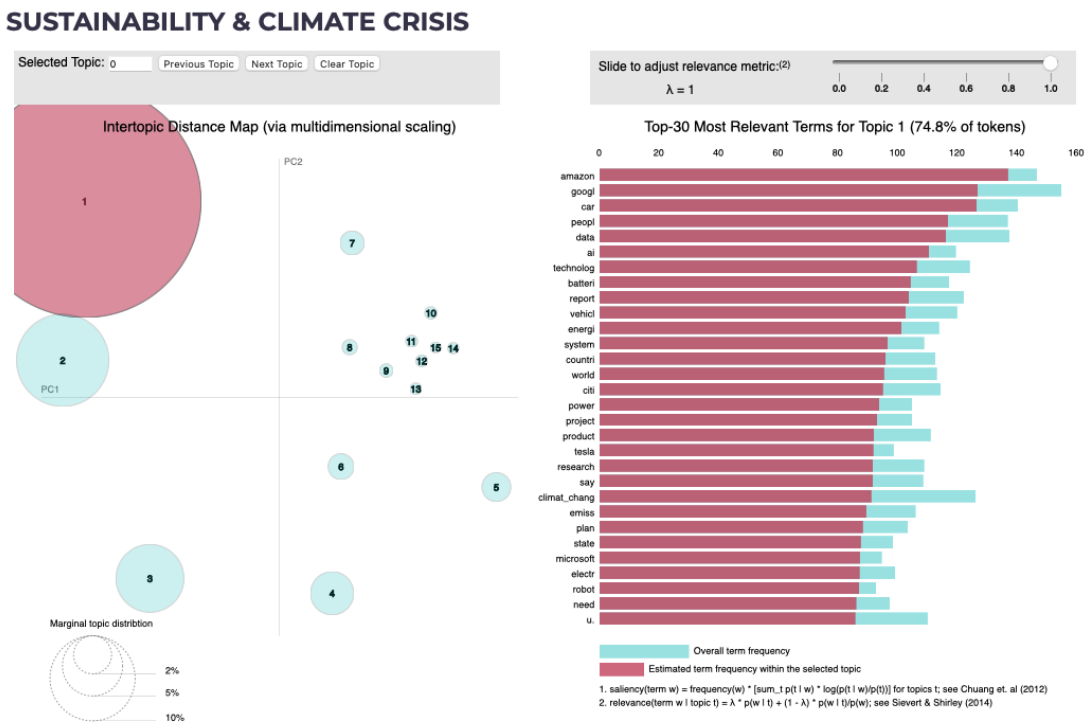
[13] https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

As part of the trend analysis pipeline, we used LDA to discover topics in selected samples of documents. Following the identification of trending terms, we created wide umbrella topics that were the basis of further analyses (co-occurrences, sentiment analysis and topic modelling). Therefore, we identified relevant articles related to the umbrella topics: the articles had to include trending keywords related to the umbrella topic. The selection of keywords combined the algorithmic and interpretive approach: first, a seed list of keywords representative of the umbrella topic was prepared. Next, all trending terms that included the root form of these seed words (e.g. root form of 'cryptocurrency' is 'cryptocurr') were identified. Finally, the list was checked manually to exclude irrelevant terms (e.g. '5G' is included in various technical terms, such as '75GB'). Following the identification of article samples, for each sample the parameters of LDA were determined with the coherence scores, the LDA was performed and the results summarised.

To conclude, LDA provides the topic-document distributions (how relevant is a topic, how frequently it is present in documents) and the word-topic distributions (how relevant is a word in a topic). This information can be well visualised using PyLDAvis[14]. In our interactive analyses, we published these interactive versions that included all discovered topics and topic-words. This format facilitates the user's own interpretation of topics. Second, in our reports we also prepared a qualitative interpretation of these results, naming the topics and presenting the most relevant words in tables.

[14] https://github.com/bmabey/pyLDAvis

Figure 7. Example for LDA



Robustness checks

We have performed extensive robustness checks for the analysis carried out during 2020 on the sample 01.2016 - 12.2019. We evaluated the identification of trending terms by splitting the dataset into two periods (first period 2016-2017, second period 2018-19) and checking whether we would have correctly predicted terms actually important in the second period using data from the first period.

Machine learning classification methods are evaluated on predicted condition versus true condition. Trend analysis is an unsupervised method – it cannot be said with certainty that a term should or should not have been selected. However, we can examine whether a term has actually gained prevalence between periods and evaluate if it has been classified as trending in the first period.

A confusion matrix tells us how many items in the dataset have been correctly or incorrectly assigned to a positive or negative class. There are four cells in the matrix: true (correct) positive, true negative, false (incorrect) positive, false negative. Based on the goal, researchers maximize particular metrics based on this matrix, such as the false positive rate (FPR, the number of false positives divided by the sum of false positives and true negatives) and true positive rate (TPR, true positives divided by the sum of true positives and false negatives). For

example, Koizumi et al. (2019)[15] analysed anomaly detection systems and minimized the FPR under a constraint that the TPR equals 1 (i.e. there are no false negatives). Essentially, their use case required that no positive is missed, while minimizing false alarms. Our primary goal is to maximise the true positive rate, identifying those terms in advance that will be deemed important in the second period. A low false positive rate is of secondary importance, as filtering in the expert evaluation stage should remove terms which would be unlikely to be relevant, for example device names.

The values of the regression coefficients have been calculated for terms based on a shorter time period: 2016-2017. Subsequently, we checked how many of the top 50, 100, 250, 500, 1000, 1500 and 2500 words predicted with our method were indeed among the "target" terms which should have been selected with full information about term frequencies from the whole period. Target terms are the terms which achieved a sufficient growth rate (1.1, 1.15, 1.25, 1.5, 2 or 3) of average frequency between the two periods, and had the highest average frequency in the second period (top 50, 100, 250, 500, 1000, 1500, 2500 – same as prediction). Based on these different parameters, we compute the TPR.

Figure 8. TPR of top 50 target terms for various growth rates and numbers of predicted words (logarithmic scale)

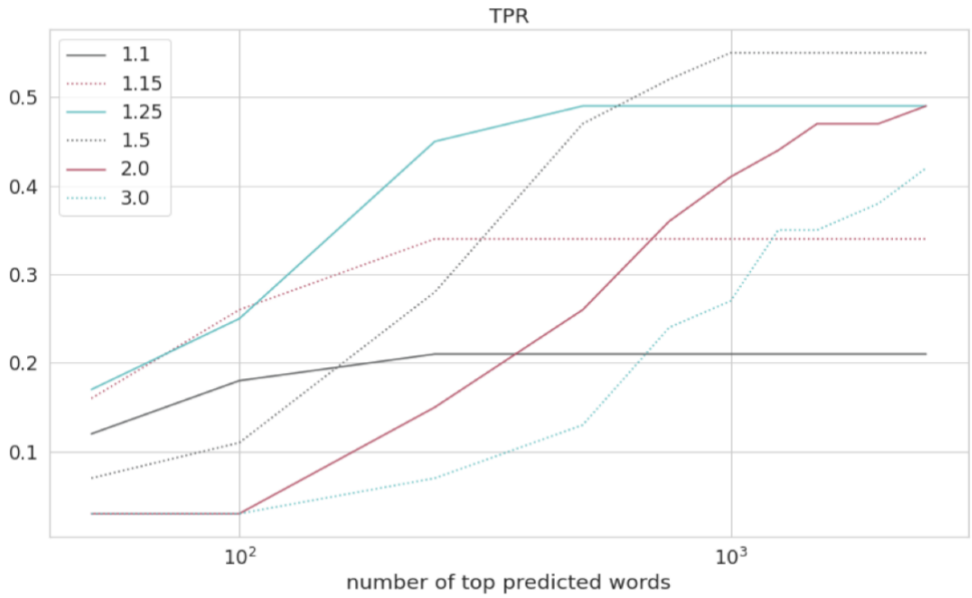


Figure 8. presents what share of the top 50 target terms were captured with our methodology based on data from the first period. The different

[15] Y. Koizumi, S. Murata, N. Harada, S. Saito, H. Uematsu, Sniper: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 915–919.

lines represent the results for the different minimum growth rates. Regardless of the precise threshold of growth, two facts crucial for the validation of our methodology remain stable. First, the results support limiting the expert analysis to the top 1000 predicted terms. While increasing the number of analysed words beyond 1000 does not bring any value in 4 out of 6 cases, the TPR value is increasing following the top 100 in almost all scenarios. Second, our methodology achieves the best results for growth rates above 1.15. TPR for a large number of words is highest for growth rate 1.5, and for low number of words – for growth rate 1.25. Our methodology does best in finding terms which would exhibit robust growth, at a cost of terms with either very high or fairly low growth rate, which are either irrelevant for stakeholders or would require thorough domain knowledge.

We also examined the TPR for terms beyond the top 50 target terms (Figure 9). TPR values are fairly high in the case of 1000 predicted terms (*top_prediction*) for target terms in the top 100 (55%) and top 250 (41%). The growth rate has been fixed at 1.5, while results for other values are available online[16].

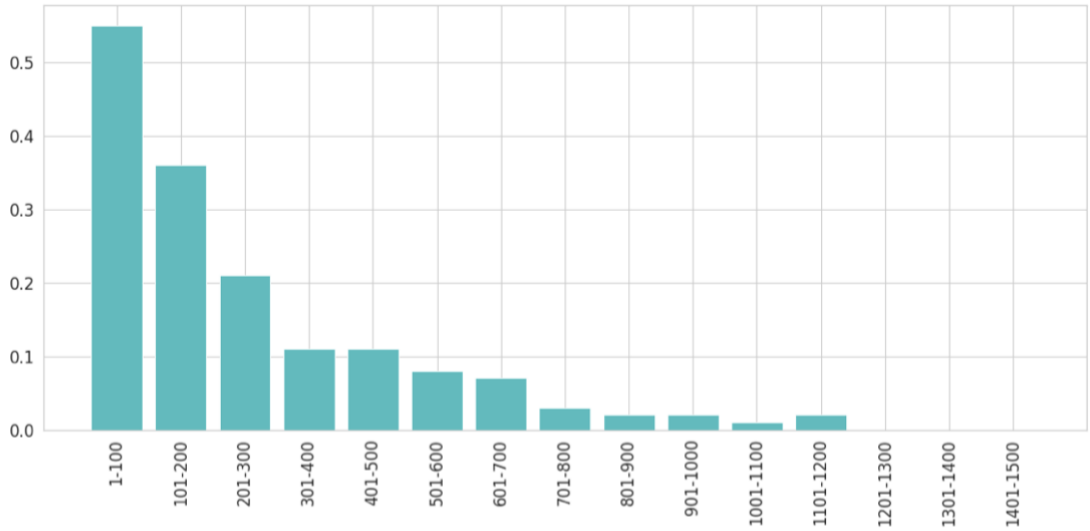
Figure 9. 20 best TPR values depending on number of predicted terms (at least 100) and number of target words

top_target	top_prediction	TPR
100	2500	0.550
100	2000	0.550
100	1500	0.550
100	1250	0.550
100	1000	0.550
100	750	0.520
250	2500	0.512
250	2000	0.492
250	1500	0.480
100	500	0.470
250	1250	0.448
500	2500	0.426
250	1000	0.408
500	2000	0.392
750	2500	0.376
250	750	0.368
500	1500	0.352
750	2000	0.332
1000	2500	0.331
500	1250	0.318

[16] https://policy.delabapps.eu/#OB1

Alternatively, we may choose a large number of target words and analyse TPR in subsets of them. Figure 10 shows the percentage of terms in a moving 100-term window which are included in the prediction set of 1000 terms. There are 15 windows, representing target terms ranked by their average frequency in 2018-19. The results support that our methodology has a high success rate of predicting growing terms, with correct prediction (TPR) not only for over 50% in the case of the top 100 target terms, but also just below 40% for terms ranked 101-200. The analysis confirms that the true positive rate is increasing with the frequency of growing terms.

Figure 10. Share of target terms (growth 1.5) in a moving window of 100 terms included in the top 1000 predicted terms



Finally, trends are not always stable over longer time periods. In order to better highlight the most recent trends, the regression analysis is calculated for shorter time periods: the last 3, 6 and 12 months[17]. The results suggest that the baseline regression performs well in capturing the most trending terms for shorter time periods (e.g. such topics as 5G, climate crisis or trade wars).

Topic mapping

The topic mapping methodology originates from our topic modelling work presented in the trend analysis. As demonstrated by our reports, LDA has been useful in indicating what kind of topics are discussed in news articles. It served as the last step in exploring the context of emerging trends: following the identification of trending keywords and selection of

[17] https://policy.delabapps.eu/#OB2

emerging umbrella topics with expert analysis, LDA complemented the insights of the co-occurrence and sentiment analyses carried out for each umbrella topic.

However, the potential of LDA is limited: the user gains an overview of topics with characteristic terms, but its use for further analysis is restricted. Therefore, we began to work on an alternative pipeline that extracts more information on topics of interest than LDA. We shifted our focus towards supporting qualitative analysis with a way to help users organise documents and read only the ones that are relevant for them. Therefore, the topic mapping pipeline not only explores keywords included in documents, but also facilitates the reading of the documents.

Our first effort combined LDA with an algorithm called t-SNE. t-SNE can be used to visualise high-dimensional data in 2D: in our case, documents are distributed in space based on their semantic similarity. t-SNE has been introduced by Van der Maaten and Hinton (2008)[18] and followed multiple attempts to find a dimensionality reduction algorithm that preserves local structure. This means that observations that are close to each other in the input high-dimensional place are also placed close to each other in the output low-dimensional space, e.g. articles about bitcoin should form a separate cloud to articles about AI, but also be close to articles about blockchain usage in smart contracts. A crucial parameter of the algorithm is perplexity, which can be understood as the number of nearest neighbors considered.
The original methodology began with LDA, exploring the wide topics in news data. Based on the distribution of topics in documents we could group the news articles based on their dominant topics, e.g. news about AI. Next, we prepared separate analyses on the major groups of articles, using both LDA and t-SNE.

We further developed this methodology in the analyses presented at ngitopics.delabapps.eu. First, our aim was to find out the most accurate methodology to group articles based on the topics discussed in them. Besides LDA, we also considered another topic model called Pachinko Allocation (PA), t-SNE and word embedding models. In order to measure and compare the accuracy of the methods, we used a reference dataset[19]. The Reuters dataset contains not only articles, but also the categories and labels for each article. Following the experiment that is presented in detail in the methodological supplement[20], we chose the t-distributed Stochastic Neighbor Embedding (t-SNE) as the

[18] Maaten, L., Hinton, G. (2008), Visualizing data using t-SNE. Journal of Machine Learning Research 9.Nov, 2579-2605.
[19] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[20] https://ngitopics.delabapps.eu/methodology.pdf

main method for the grouping of articles. While the results confirmed that there is no single state-of-the-art method which outperforms all other methods across all metrics, the methodology based on t-SNE has proven to be robust and consistent.

The pipeline

The philosophy behind topic mapping is different from trend analysis: while trend analysis focuses on finding emerging keywords and exploring them, topic mapping captures all areas of discussion. Therefore, the first step in the analysis is the definition of wide areas the researcher is interested in, e.g. issues related to privacy or misinformation. In our analyses based on topic mapping, we focused on six wide umbrella topics selected together with our partners from the project. Next, we compiled a dataset of online documents corresponding to the six topics. The details of the data collection can be found in the report[21].
Following the results of the Reuters case study, we used t-SNE to map articles in two dimensions: articles close to each other on the map cover the same issue. The maps also contain various colours that show the wide clusters of documents. These clusters were assigned with the use of a clustering algorithm (Gaussian mixtures), hence the colours highlight documents that belong together based on their location on the map.

Next, using qualitative analysis, we have examined the text and titles of articles to select the most relevant clusters and assigned the tag descriptions. These topics are highlighted on the maps: the arrow is indicating the subject and the location of the article cluster. Finally, we also prepared deep dives for selected issues based on the maps to showcase the potential of the methodology. The deep dives are based on the articles presented in the article clusters. We focused on areas, where the results highlighted not only the underlying challenges, but also policy and technological solutions. These deep dives are interpretive: other researchers may choose other clusters and topics for analysis. As all interactive maps are publicly available[22], all users can verify the clusters and explore the articles.

The global nature of the challenges means that a wide range of discussions and debates take place across the world in different languages. Therefore, by restricting the analysis only to English, we might be losing valuable country-specific information on the analysed topics. For this reason, we also prepared a complementary analysis for articles in German, Polish, Spanish and Portuguese[23].

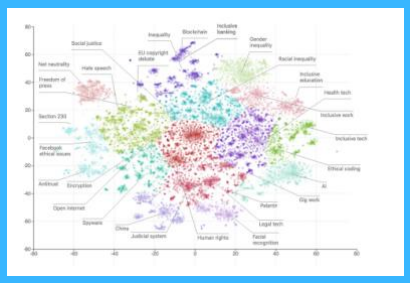
[21] ngitopics.delabapps.eu/report.pdf
[22] ngitopics.delabapps.eu
[23] ngitopics.delabapps.eu/report_multilanguage.pdf

Figure 11. Topic mapping pipeline

Agenda setting

define the umbrella topics, keywords and identify social media platforms

01



02

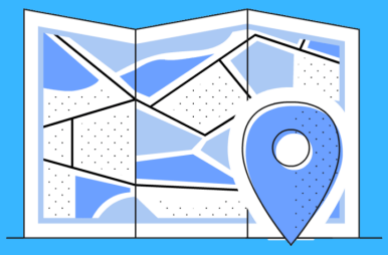
Algorithms

Create article "maps" using t-SNE and clustering algorithms

Interpretation #1

Select topic keywords from a list of characteristic terms created with text-mining

03



04

Interpretation #2

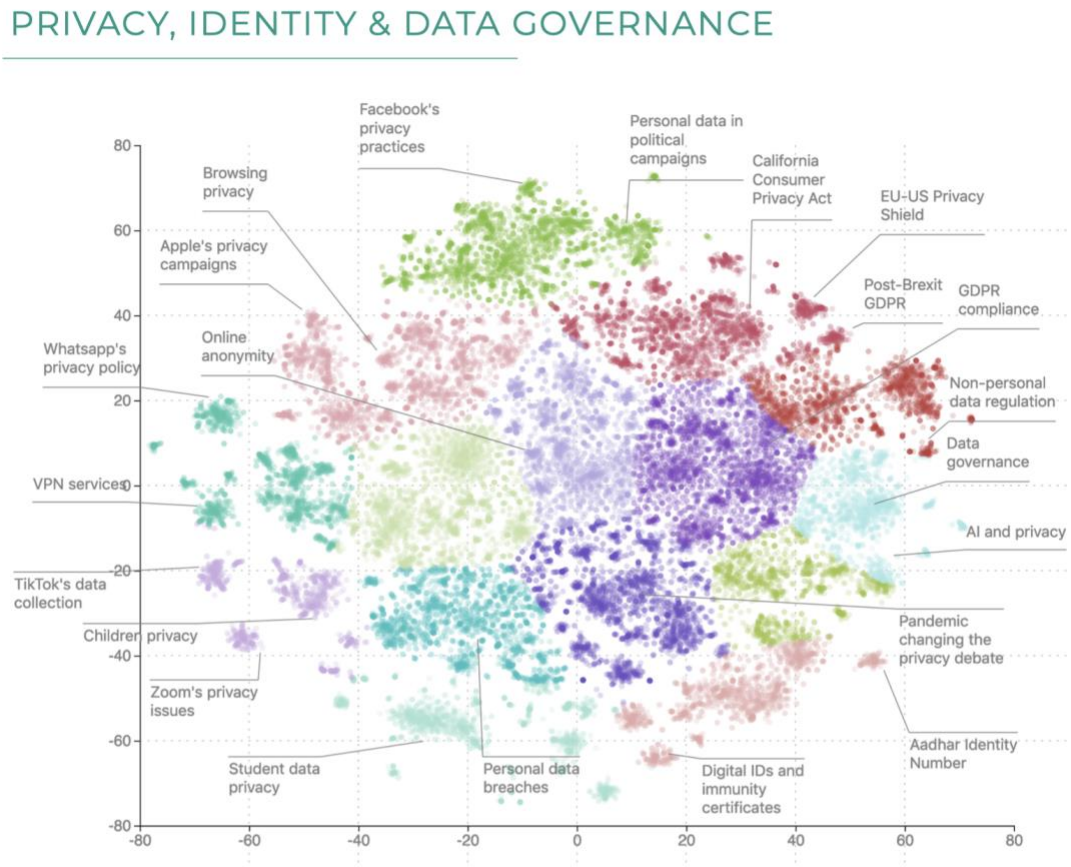
Tag maps based on the articles that form the small groups.

Interpretation #3

Find challenges and solutions, shown in the case study

05

Figure 12. Example of topic mapping



Robustness check

The robustness of the selected methodology was extensively presented in the methodology paper . Additionally, we provided an alternative pipeline, where t-SNE is combined with HDBSCAN clustering. In the paper[24], we have demonstrated that HDBSCAN has many advantages in case of a completely automated analysis, while Gaussian mixtures is more suited for qualitative analysis of results.

[24] https://ngitopics.delabapps.eu/methodology.pdf

Datasets

Online News dataset

This dataset has been the basis of the trend analysis. The dataset contains news articles with accompanying metadata collected for a period of 5 years and 4 months between 2016-01-01 and 2021-04-30. We provide all details on the dataset in this section: the last detailed report on trend analysis was prepared in 2020, hence this final report presents the final, most complete version of the dataset.

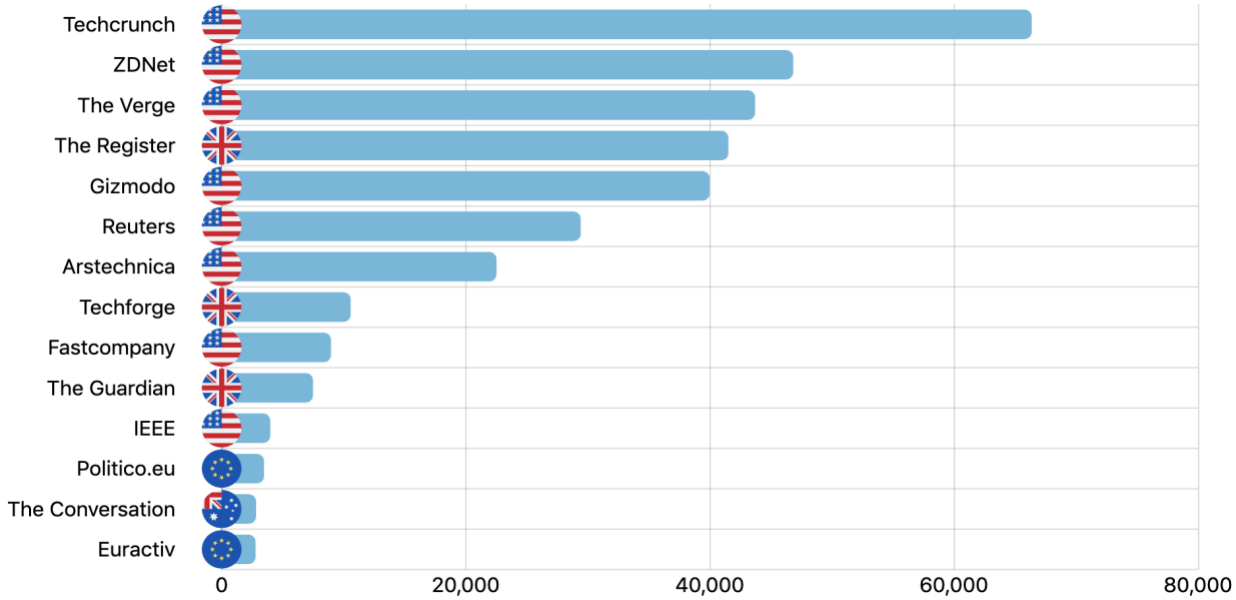
The datasets with raw text data can be prepared using the web-scraping scripts published on Gitlab[25]. A separate script has been prepared for every news portal. While most of the scrapers are based on the web-automation framework Selenium Webdriver, in the case of the Guardian, API access is provided, enabling data collection directly from the publisher.

Table 1. Collected data: online news articles

Metadata	Description
Title	Article title
Date	Publication time of article (YYYY-MM-DD)
Abstract	Short description of the article (1 sentence) if applicable
Link	Article link
Text	The entire text of the article is in merged form (paragraphs are marked by \n characters)

[25] https://gitlab.com/enginehouse/ngi-forward-d1-6

Figure 13. Number of media articles per source and country of origin



1. Ars Technica

Ars Technica is a US and UK based tech website, covering news and opinions in technology, science, politics, and society. This data-set contains news articles scraped from the Ars Technica website (https://arstechnica.com/). The dataset consists of the article text and various metadata. Articles from the following news categories have been collected:

- Information-Technology
- Gadgets
- Science
- Tech-Policy
- Cars

Number of articles collected: 22 465

2. Euractiv

EURACTIV is an independent pan-European media network specialised in EU policies. This data-set contains news articles scraped from the Euractiv website (https://www.euractiv.com/). The dataset consists of the article text and various metadata. Articles from the Digital category have been collected.

Number of articles collected: 2 726

3. Fastcompany

Fast Company is a US-based monthly magazine that focuses on the latest developments in technology, business, and design. This data-set contains news articles scraped from Fast Company (https://www.fastcompany.com). The dataset consists of the article text and various metadata. Articles from the category “Technology” were retrieved.

Number of articles collected: 8 903

4. Gizmodo

Gizmodo is a US-based technology blog network covering topics like technology, science, science-fiction,

pop culture, design and politics. This data-set contains news articles scraped from Gizmodo (http://www.gizmodo.co.uk/). The dataset consists of the article text and various metadata. Articles from the following categories have been downloaded (all except Gizmodo's subsite IO9, which is mostly concerned with covering movies, sci-fi literature and popular culture, and therefore not very relevant to our analysis)

- Sploid
- Paleofuture
- Science
- Field Guide
- Design

Number of collected articles: 39 934

5. IEEE Spectrum

IEEE Spectrum is the official magazine of the Institute of Electrical and Electronics Engineers, the world's largest professional organisation for engineers, with over 400.000. IEEE Spectrum covers technology, engineering and science news. This data-set contains news articles scraped from IEEE Spectrum (https://spectrum.ieee.org/). The dataset consists of the article text and various metadata. Articles from the following categories have been collected:

- Automaton
- Tech-talk
- The human-os
- Riskfactor
- View from the valley
- Nanoclast
- Cars that think

Number of collected articles: 3927

6. Politico Europe

POLITICO is a global nonpartisan politics and policy news organization, launched in Europe in April 2015. POLITICO Europe is a joint-venture between POLITICO LLC, based in the USA and Axel Springer. This data-set contains news articles scraped from the Politico Europe website (https://www.politico.eu/). The dataset consists of the article text and various metadata. Articles from the Technology and Data and Digitization sections have been collected.

Number of collected articles: 3 424

7. Reuters

Reuters is one of the world's largest news agencies, based in London, UK. It brings together the latest, fact-based reporting on a variety of topics from across the globe. This data-set contains news articles scraped from the Reuters website (https://https://www.reuters.com). The dataset consists of the article text and various metadata. Articles from the Technology section have been collected (https://www.reuters.com/news/archive/technologyNews).

Number of collected articles: 29 363

8. TechCrunch

TechCrunch is an American online publisher focusing on the tech industry. The company specifically reports on the business related to tech, technology news, analysis of emerging trends in tech, and

profiling of new tech businesses and products. This data-set contains news articles scraped from the TechCrunch website (https://techcrunch.com/). The dataset consists of the article text and various metadata. Articles from the following sections have been collected:

- Startups
- Apps
- Gadgets

Number of collected articles: 66 317

9. Techforge

TechForge is an independent publishing company on tech related publications, home to over 15 sub-brands on specific technologies such as IoT, Virtual Reality and Cloud Computing. This data-set contains news articles scraped from a portfolio of technology news websites published by Techforge (https://www.techforge.pub/). The dataset consists of the article text and various metadata. Articles from the following Techforge websites have been downloaded:

- http://www.virtualreality-news.net
- https://www.cloudcomputing-news.net/
- https://www.telecomstechnews.com/
- https://www.developer-tech.com/
- https://www.enterprise-cio.com/
- https://www.iottechnews.com/
- https://www.marketingtechnews.net/

Number of collected articles: 10 522

10. The Conversation

The Conversation is an independent, not-for-profit media outlet that brings together the most interesting academic research and turns it into accessible outputs for a more general public. This data-set contains news articles scraped from the website The Conversation (https://theconversation.com). The dataset consists of the article text and various metadata. Articles from the Technology section have been collected (https://theconversation.com/uk/technology/articles).

Number of collected articles: 2 764

11. The Guardian

The Guardian is one of the UK's largest daily newspapers, based in London. This data-set contains news articles scraped from The Guardian (https://www.theguardian.com). The dataset consists of the article text and various metadata. Articles have been collected from the Technology category (https://www.theguardian.com/uk/technology).

Number of collected articles: 7 432

12. The Register

The Register is a British technology and opinion website. This data-set contains news articles scraped from The Register (https://www.theregister.com). The dataset consists of the article text and various metadata. Articles have been collected from all categories, using the Archive section of the website

(https://www.theregister.co.uk/Archive).

Number of collected articles: 41 474

13. The Verge

The Verge was founded in 2011 in partnership with Vox Media, and covers the intersection of technology, science, art, and culture. This data-set contains news articles scraped from Verge website (https://www.theverge.com/). The dataset consists of the article text and various metadata. Articles have been collected from all tech subsections.

Number of collected articles: 43 653

14. ZDNet

ZDNet is a US-based technology publication, with a focus on general technology news. This data-set contains news articles scraped from ZDNet (https://www.zdnet.com/). The dataset consists of the article text and various metadata. Articles have been collected from the following categories:

- Artificial intelligence
- banking
- data centers
- data management
- developer
- digital transformation
- e-commerce
- enterprise software
- EU
- Future of Work
- Google
- Government
- Great debate
- Innovation
- Internet of things
- IT priorities
- legal
- mastering business analytics
- networking
- open source
- reimagining the enterprise
- robotics
- security
- smart cities
- social enterprise
- start-ups
- tech industry
- virtual reality
- 3d printing

Number of collected articles: 46 780

Social media articles dataset

Our second major dataset followed a different logic than the News Media dataset: instead of collecting all articles published by a news portal, articles shared on social media on particular topics were collected. We have selected Twitter, Reddit and Hacker News as our data sources. In the selection process we have considered the following criteria:

- Tech-oriented user base
- Diverse audience (Twitter mostly representing tech mainstream, while Reddit's and Hacker News' userbase is tech-savvy)
- Data availability (API or BigQuery storage)

In the first step, we have collected posts containing selected keywords and links to media articles. In this process we have utilized - the official Twitter API, Reddit's Pushshift API and Hacker News BigQuery

database. In the case of Twitter, the sample sizes required reduction for further analysis. Therefore, to focus on the most influential content, posts that were retweeted less than three times were dropped from the analysis. Next, we have extracted article text and metadata with the use of the Python package Newspaper3k[26].
Moreover, we prepared the data collection not only for English-language social media posts, but also for posts in German, Polish, Portuguese and Spanish. Details on the datasets are provided in the deliverable reports.

Academic working papers dataset

In the case of working papers, works from two repositories are collected: arXiv (STEM sciences) and SSRN (social sciences) for the period 01.2016 - 06.2020.

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. The data-set consists of working papers acquired via ArXiv's API (https://arxiv.org/help/api/index). The data-set consists of the working papers' abstracts and various metadata. Working papers (called e-prints by ArXiv) have been collected from the Computer Science discipline (all CS categories).

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-set is comprised of the working papers' abstracts and various metadata. Working papers have been collected from two broad categories (called Research networks by SSRN administrators): 1. Information Systems & eBusiness, 2. Innovation.

The overall number of the analysed working papers on arXiv reaches 155 579, while working papers on social sciences include over 25 151.

[26] https://github.com/codelucas/newspaper

All resources created during the project

Reports with accompanying presentations and datasets

Visualisations of key emerging technologies and social issues I. (June 2019)

The report summarises the results of the first iteration of trend analysis. The report is based on the online news dataset and on the working paper dataset for the period 2016.01-2019-04.
The datasets with raw results are available at Zenodo:

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2019). Keyword frequencies in popular tech media (01.2016-04.2019) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3243926

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2019). Co-occurrences of trending keywords in popular tech media (01.2016-04.2019) (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3243949

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2019). Sentiment analysis of tech media articles using VADER package and co-occurrence analysis (01.2016-04.2019) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3243961

The identified umbrella topics, chosen from the most trending technologies and issues are the following:

- 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

The report is available at Zenodo: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, & Maciej Wilamowski. (2021). Visualisations of key emerging technologies and social issues I. Zenodo. https://doi.org/10.5281/zenodo.5796225

The accompanying interactive presentation is available at: https://fwd2019.delabapps.eu/

Intermediary topic modelling analysis results (August 2019)

The report provides our preliminary analysis for topic mapping. The analysis is based on the same news media dataset as the first iteration of trend analysis (2016.01-2019-04). The presented methodology used LDA and t-SNE to map latent topics in the news dataset. 17 wide umbrella topics were identified, and 5 were selected for further analysis:

1. AI and Robots
2. Policy (Social media crisis, Privacy and 5G)
3. Media
4. Business
5. Cybersecurity

The report is available with interactive visualisations at: https://fwd.delabapps.eu/topic_modelling.html

Visualisations of key emerging technologies and social issues II. (August 2020)

The report summarises the results of the second iteration of trend analysis and an additional deep dive focused on the COVID-19 pandemic. Therefore, two separate trend analyses are presented: one for the period 2016.01-2019.12, and the second for the period 2020.01-2020.06. The COVID-19 analysis includes further explorations on open-source Github projects and about Reddit discussions.

The datasets with raw results are available at Zenodo:

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2020). Keyword frequencies in popular tech media (01.2016-12.2019) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3942060

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2021). Keyword frequencies in popular tech media during the COVID-19 pandemic (01.2020-06.2020) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5788201

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2020). Co-occurrences of trending keywords in popular tech media (01.2016-12.2019) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3942104

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2021). Co-occurrences of trending keywords in popular tech media during the COVID-19 pandemic (01.2020-06.2020) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5788325

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2020). Sentiment analysis of tech media articles using VADER package and co-occurrence analysis (01.2016-12.2019) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3942124

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2021). Sentiment analysis of tech media articles using VADER package and co-occurrence analysis during the COVID-19 pandemic (01.2020-06.2020) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5788353

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2021). Keyword frequencies in arXiv and SSRN working papers [Data set]. Zenodo.

https://doi.org/10.5281/zenodo.5789536

The identified umbrella topics, chosen from the most trending technologies and issues are the following:

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

The report is available at Zenodo: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, & Maciej Wilamowski. (2021). Visualisations of key emerging technologies and social issues II. Zenodo. https://doi.org/10.5281/zenodo.5796283

The accompanying interactive presentation is available at: https://fwd2020.delabapps.eu/
The COVID-19 presentation is available at: https://covid.delabapps.eu

Final topic modelling analysis results (August 2021)

This report introduced the final topic mapping methodology based on t-SNE. The analysis has been prepared on the basis of the social media articles dataset that covered six wide umbrella topics:

- Environment, Sustainability & Resilience
- Decentralising Power & Building Alternatives
- Public Space & Sociality
- Privacy, Identity & Data Governance
- Trustworthy Information Flows, Cybersecurity & Democracy
- Access, Inclusion & Justice

The datasets with the titles and links of the collected articles are available at Zenodo:

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2021). Social media articles dataset - article titles and links [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5789701

There are three major outputs: the methodology paper “Identifying topics in unlabeled documents: a methodological guide”, a report summarizing the results based on the social media articles dataset “Towards a Human-Centric Internet: Challenges and Solutions” and a website presenting the interactive visualisations and hosting the reports available at https://ngitopics.delabapps.eu.
The reports are also available at Zenodo: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, Maciej Wilamowski, & Katarzyna Śledziewska. (2021). Final topic modelling analysis - Towards a Human-Centric Internet: Challenges and Solutions. Mapping Key Tech And Policy Topics With Text-mining.

Zenodo. https://doi.org/10.5281/zenodo.5796491

Visualisations of key emerging technologies and social issues III (November 2021)

The report titled "Towards a Human-Centric Internet: A multi-language analysis of key tech and policy topics" continued the work of the previous topic modelling analysis", using the same topic mapping methodology for articles in various languages: German, Polish, Spanish and Portuguese. The dataset has been published in the same Zenodo repository:

Gyódi, Kristóf, Nawaro, Łukasz, & Paliński, Michał. (2021). Social media articles dataset - article titles and links [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5789701

Similarly, the results are published in interactive form at https://ngitopics.delabapps.eu.

The report is also available at Zenodo: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński, Maciej Wilamowski, & Katarzyna Śledziewska. (2021). Visualisations of key emerging technologies and social issues III: Towards a Human-Centric Internet: A Multi-Language Analysis of Key Tech and Policy Topics. Zenodo. https://doi.org/10.5281/zenodo.5796501

Finally, we also prepared the third iteration of the trend analysis on the news media dataset for the entire time period of the data collection: 01.2016 - 04.2021. The results are summarized in the interactive presentation: https://fwd.delabapps.eu

To keep our final analysis consistent with topic mapping, we organized the results matching the six umbrella topics from the previous analyses. However, the results did not support that the relevance of Public Space & Sociality topic is increasing (e.g. smart city has not been identified as a trending term), hence it was not included among the deep dives. Naturally, terms related to the pandemic have been heavily represented in the results, therefore we added the Coronavirus pandemic & Consequences umbrella topic to the analysis:

- Environment, Sustainability & Resilience
- Decentralising Power & Building Alternatives
- Privacy, Identity & Data Governance
- Trustworthy Information Flows, Cybersecurity & Democracy
- Access, Inclusion & Justice
- Coronavirus pandemic & Consequences

The raw results are available in:

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2021). Keyword frequencies in popular tech media (01.2016-04.2021) [Data set]. Zenodo.

https://doi.org/10.5281/zenodo.5788431

Kristóf Gyódi, Nawaro, Łukasz, & Paliński, Michał. (2021). Co-occurrences of trending keywords in popular tech media (01.2016-04.2021) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5788439

Kristóf Gyódi, Łukasz Nawaro, & Michał Paliński. (2021). Sentiment analysis of tech media articles using VADER package and co-occurrence analysis (01.2016-04.2021) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5789501

Codes and tutorials

The documented codes and tutorials introducing our text-mining methods are published on Gitlab: https://gitlab.com/enginehouse/ngi-forward-d1-6. The programming codes have been prepared in the Python programming language. The codes and notebooks are readable from the web browser, therefore anyone interested can get familiar with the tutorials.

The essential sections are the following:

- Readme: summary of the repository structure
- Scrapers: folder containing the web-scraping scripts, including the Twitter search queries
- Requirements.txt: File containing the required Python packages to run the analyses. The tutorials include the bulk installation of all packages
- Imposing_functions.py: Python script containing functions used in the analyses
- Main_settings.py sets paths and major parameters

Moreover, for better understanding and greater impact, 6 more notebooks are published that explain the analyses and guide users on how to replicate them. The notebooks are provided in the Jupyter Notebook format that facilitates the exploration of codes in smaller chunks. Therefore, the tutorials enable users to gain hands-on experience with minimal prior programming knowledge (installation of Python and Jupyter Notebook).

The following tutorials are published:

0_text_transformation_guide.ipynb: the documentation for the functions, explaining how text transformation functions work in the NGI Forward project. This is essential for data-scientists as it provides the framework for our analyses. However, less advanced users can skip this tutorial as it is not required to change any of the functions.

1_initial_transformation.ipynb: downloads the articles using the Guardian API and prepares the data in the format suitable for the text mining analysis conducted in the tutorials.

A_topic_identification.ipynb: this is the first tutorial with text-mining analyses. The notebook will guide the user step-by-step through the trend analysis pipeline we have developed for identifying trending technologies and related social issues in tech media. The tutorial covers:

- identification of trending keywords using OLS regression
- analysis of keywords co-occurrences
- sentiment of trending topics using VADER package

B1_LDA.ipynb: presents how to prepare and visualise the results of the topic modelling analysis using Latent Dirichlet Allocation. The user is also supported with selecting optimal parameters.

B2_t-SNE.ipynb: shows how to combine the LDA analysis with the t-SNE algorithm in order to map clusters of articles.

C_t-SNE.ipynb: presents the final topic mapping methodology.

Additionally, we also prepared a tutorial for collecting data from HackerNews: Michał Paliński. (2021). HackerNews analysis with BigQuery, Kaggle tutorial: https://www.kaggle.com/michapaliski/hackernews-analysis-with-bigquery

The tutorial explains all steps of data collection from HackerNews with Google's BigQuery. Following the tutorial, the user can set up from scratch the necessary work environment and run search queries on HackerNews data. Moreover, the tutorial shows how to download the entire discussions belonging to a post, even comments to comments.

Academic papers

Our paper titled: "Informing policy with text mining: technological change and social challenges" has received a major revision decision at the journal Quality and Quantity, we are in the process of reviewing the paper. The paper is summarizing the trend analysis methodology on the basis of the news media dataset for the period 01.2016-12.2019. The various supplementary materials of the paper are summarized online: https://policy.delabapps.eu

Blog posts

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński & Maciej Wilamowski. (2019). Data Science Tools for Technology Mapping, NGI Forward, https://research.ngi.eu/data-science-tools-for-technology-mapping/

The post provides a brief introduction to trend analysis with a focus on AI and ML.

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński. (2020). Making sense of the COVID-19 information maze with text-mining, NGI Forward, https://research.ngi.eu/making-sense-of-the-covid-19-information-maze-with-text-mining/

This blog post presents a short summary of the COVID-19 analysis.

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński & Maciej Wilamowski. (2019). Mapping the tech world with topic modelling, Towards Data Science, available: https://towardsdatascience.com/mapping-the-tech-world-with-topic-modelling-bfc3c40af507

Kristóf Gyódi, Łukasz Nawaro, Michał Paliński & Maciej Wilamowski. (2019). Mapping the tech world with t-sne, Towards Data Science, available: https://towardsdatascience.com/mapping-the-tech-world-with-t-sne-7be8e1703137

These two blog posts present the short summary of the report intermediary topic modelling analysis results. The first post introduces LDA and explains how to interpret the interactive visualisations. The second writing continues the analysis with t-sne.

Kristóf Gyódi. (2021). Discussing privacy at HackerNews: An explorative text-mining analysis, Towards Data Science, available: https://towardsdatascience.com/discussing-privacy-at-hacker-news-an-explorative-text-mining-analysis-f94c62802df9

This blog post presents an analysis of HackerNews discussions about privacy using sentiment analysis. The basis of the exercise is the HackerNews analysis with BigQuery tutorial. The post explains the potential of text-mining methods in extracting insights from social media posts and comments.

NEXT GENERATION INTERNET
INTERNET OF HUMANS

NGI FORWARD

This report was created by DELab for NGI Forward, part of the Next Generation Internet initiative, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°825652

Websites:
https://www.ngi.eu
https://ngitopics.delabapps.eu
https://research.ngi.eu
https://fwd.delabapps.eu
https://www.delab.uw.edu.pl

Twitter
https://twitter.com/ngi4eu
https://twitter.com/ngiforward
https://twitter.com/delabuw