D1.5. Intermediary topic modelling analysis results

Work package	WP1: Topic identification
Task	1.3 Topic co-occurrence analysis: where do technology and social issues meet?
Due date	01.11.2019
Submission date	31.10.2019
Deliverable lead	DELab
Dissemination level	Public
Nature	Report
Authors	DELab: Kristóf Gyódi, Łukasz Nawaro, Michał Paliński
Version	1.2
Reviewers	Katja Bego, Heshani Jayaratne
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 Nº825652**

Mapping the tech world using text mining methods

Head to https://fwd.delabapps.eu/topic_modelling.html to view a full interactive version of the deliverable D1.5

Executive summary

This study presents an innovative methodology for analysing technology news using various text mining methods. News articles provide a rich source of information to track promising emerging technologies, relevant social challenges or policy issues. Our goal is to support the Next Generation Internet initiative by providing data-science tools to map and analyse the developments of the tech word.

Based on more than 200 000 articles from major media outlets, we are going to identify widely discussed topics, focusing on emerging technologies and policy issues and dive deeper in selected areas and highlight key focal points of recent developments.

To meet these goals, a number of machine learning techniques are combined. The major steps can be summarised as follows:

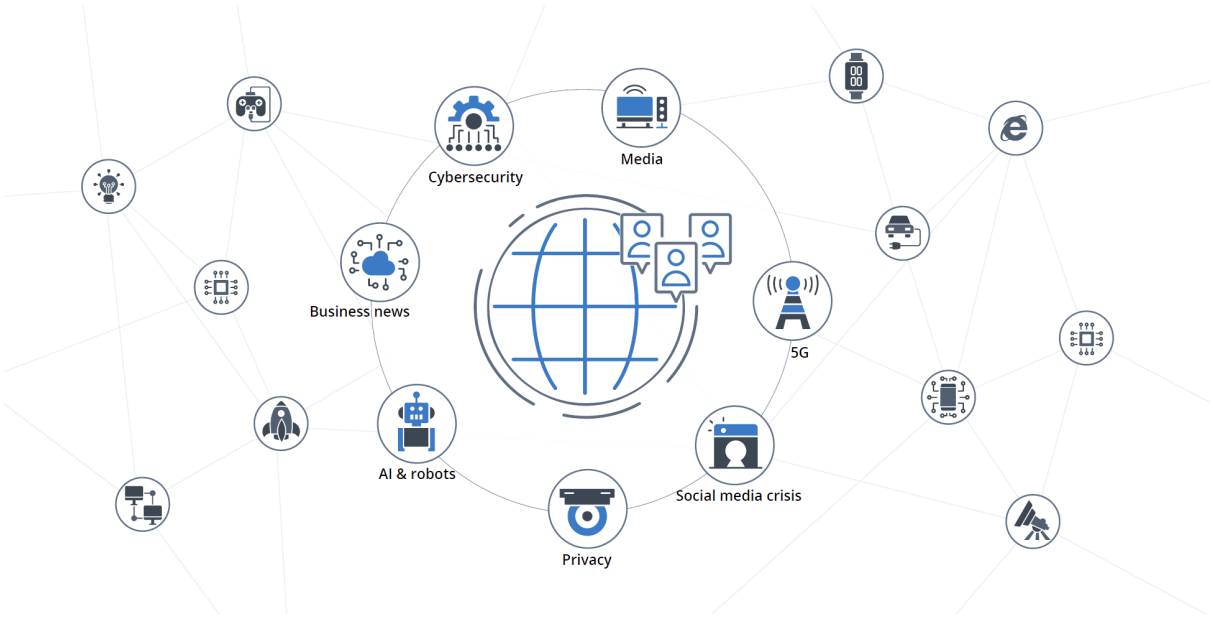
- 17 general umbrella topics are explored
- 5 topics are selected for further analysis
- Deep dives are presented with 2D interactive maps

More specifically, the topics selected for the deep dives are:

1. AI and Robots
2. Policy (sums up 3 relevant areas)
3. Media
4. Business
5. Cybersecurity

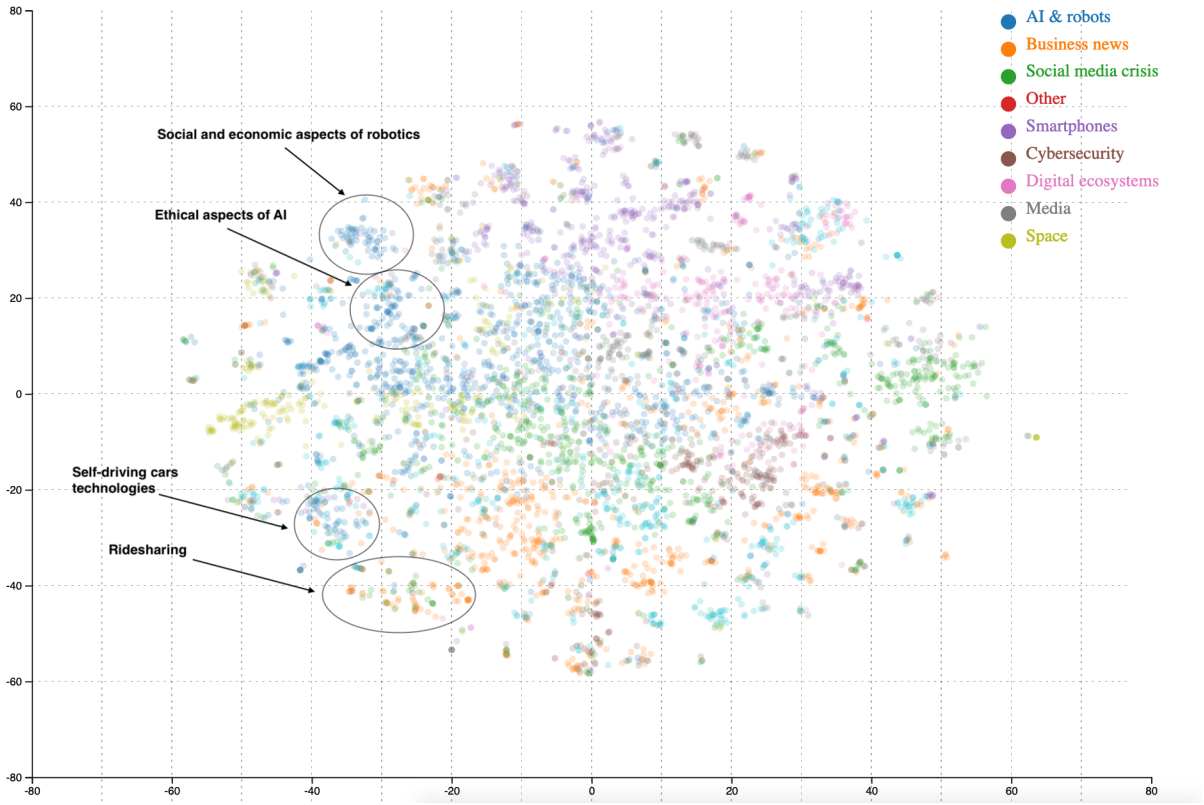
With the *Policy* topic grouping together 3 areas: *Social media crisis*, *Privacy* and *5G*.

Wide areas selected for deep-dive analyses



The 17 umbrella topics are identified using the topic modelling technique *Latent Dirichlet Allocation*. Besides the topics selected for deep dives, such areas are highlighted as *Smartphones*, *CPU and other hardware*, *Digital ecosystems* or *Space*.

Next, various maps are created based on the t-SNE algorithm. The example below presents the news stories in two-dimensions: articles that report on the same subject are clustered together. We demonstrate that this technique is highly useful to discover more narrow, domain-specific areas within the umbrella topics. Moreover, the distance between clusters is also meaningful, enabling the analysis of relationships between topics as well.



As an example, within the *AI and robots* topic, the map reveals groups of articles focused on such issues as:

- social and economic challenges of robotic technologies (e.g. job automation),
- ethical aspects of AI (e.g. military AI),
- AI advancements in the field of self-driving cars.

It is also visible that articles on social and ethical issues are closer to each other, while articles on AI in self-driving cars are placed near business news on ride-sharing apps. It shows that our methodology is efficient in decreasing the complexity of text data, enabling to analyse and map topics.

All maps are interactive, inviting users to explore the headline of articles.

Example articles

Social and economic aspects of robotics

- *Robots do destroy jobs and lower wages*
- *Zume's robot pizzeria could be the future of workplace automation*
- *RoboCop is real and could be patrolling a mall near you*

Ethical aspects of AI

- *Rise of the racists robots - how AI is learning all our worst impulses*
- *You should find out what's going on in that neural network. Y'know they're cheating now?*
- *Can you tell the difference between a real face and an AI-generated fake?*

Self-driving cars technologies

- *Where's the lane? Self-driving cars confused by shabby U.S. roadways*
- *Jaguar's new concept has a steering wheel you talk to and take with you*
- *Owl camera watches over your ride from the inside*

The presented methodology provides intuitive, easily understandable results. To enhance the exploration of results, the study is presented as an interactive guide. This report has been designed with different readers in mind, offering various journeys. To analyse and understand the results, it is sufficient to read the introduction and results sections. We also prepared a guide briefly explaining various text-mining methods for anyone interested. Finally, detailed description of methods are included for proper reproducibility of the study in the methods section.

The theoretical background and a clear outline of the methodology used in this deliverable, alongside an extensive description of the the results introduced above are available in the interactive version of the deliverable.

Head to https://fwd.delabapps.eu/topic_modelling.html to view a full interactive version of the deliverable D1.5