

Scientific publishing and collaboration in ecosystems

**Analysis of Finnish publications in
Flagship-related topics**

Valeria Caras, Yrjö Leino and Aino Ropponen



ACADEMY OF FINLAND

Sisällys

| | |
|--|-----------|
| Tiivistelmä..... | 3 |
| Sammanfattning | 3 |
| Summary | 4 |
| 1. Introduction | 4 |
| 2. Data and methods | 5 |
| 2.1. Data description | 5 |
| 2.2. Modeling | 7 |
| 3. Bibliometric analysis..... | 9 |
| 3.1. Number of publications | 10 |
| 3.2. Collaboration | 11 |
| 3.3. Subject fields..... | 12 |
| 3.4. Scientific impact..... | 18 |
| 3.5. Collaboration networks | 19 |
| 4. Conclusions | 21 |
| References | 23 |
| Appendix A. Keywords and publications by Flaghip topic | 25 |
| Appendix B. Literature review of machine learning methods for textual data..... | 26 |
| Appendix C. Abbreviations for organizations..... | 28 |

ISBN 978-951-715-938-8

Tiivistelmä

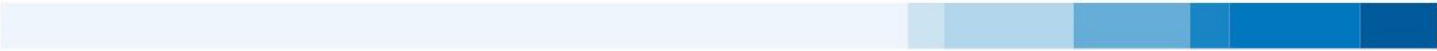
Suomen Akatemia käynnisti lippulaivaohjelman vuonna 2017 tukemaan korkeatasoista tutkimusta ja edistämään tutkimuksen yhteiskunnallista vaikuttavuutta. Tässä raportissa tarkastellaan lippulaivoihin liittyvien tutkimusaiheiden kehitystä analysoimalla Web of Science (WoS) Core Collection -julkaisuja vuosilta 2005–2008 ja 2015–2018. Lippulaivojen raportoimia julkaisuja käytettiin aineistona, jolla koulutettiin koneoppimiseen perustuvaa mallia (Supervised Machine Learning model – Random Forest Binary classifier). Näin yhteensä 2 155 julkaisulla (80 % koulutusmalleja varten ja 20 % testausta varten) tunnistettiin WoS-tietokannasta 7 579 tutkimusaiheisiin liittyvää julkaisua vuosilta 2005–2008 ja 11 381 julkaisua vuosilta 2015–2018.

Bibliometrisessä analyysissä tutkittiin suomalaisten tutkimusorganisaatioiden ja kansainvälisten kumppaneiden välistä yhteistyötä ja tieteellistä vaikuttavuutta. Tulokset osoittavat, että lippulaiva-aiheissa tehdään tieteellisesti vaikuttavaa tutkimusta. Lippulaivateemojen julkaisujen top 10 -indeksi, joka kuvaa eniten viitattuun 10 prosenttiin kuuluvien julkaisujen osuutta, on korkeampi kuin maailman keskiarvo molemmilla ajanjaksoilla. Lisäksi lippulaivojen teemoissa tehtiin tiivistä kansallista ja kansainvälistä yhteistyötä julkaisutoiminnassa. Suomalaisten tutkimusorganisaatioiden yhteisjulkaisujen määrä erityisesti ulkomaisten kumppaneiden kanssa kasvoi huomattavasti tarkastelujakson aikana.

Sammanfattning

Finlands Akademi inledde flaggskeppsprogrammet år 2017 för att stödja högklassig forskning och främja forskningens samhälleliga genomslag. I denna rapport granskas utvecklingen av de forskningsteman som flaggskeppen behandlar genom en analys av publikationer i databasen Web of Science (WoS) Core Collection åren 2005–2008 och 2015–2018. De publikationer som flaggskeppen hade rapporterat användes som underlag för att skola en maskininlärningsmodell (Supervised Machine Learning model – Random Forest Binary Classifier). Med sammanlagt 2 155 publikationer (80 % för modellerna och 20 % för testning) identifierade man från WoS-databasen sedan 7 579 publikationer från åren 2005–2008 och 11 381 publikationer från åren 2015–2018 som var förknippade med forskningsteman.

I den bibliometriska analysen undersöktes samarbetet och den vetenskapliga genomslagskraften mellan finländska forskningsorganisationer och internationella partner. Resultaten visar att det bedrivs vetenskapligt effektiv forskning inom flaggskeppsteman. Publikationer inom flaggskeppsteman hade ett topp 10-index (beskriver andelen publikationer som hör till den 10 procents andel publikationer som fått flest citeringar) som är högre än genomsnittet i världen under båda perioderna. Dessutom har forskarna inom flaggskeppens teman haft ett nära nationellt och internationellt samarbete



inom publikationsverksamheten. Antalet sampublikationer vid finländska forskningsorganisationer, särskilt med utländska partner, ökade avsevärt under granskningsperioden.

Summary

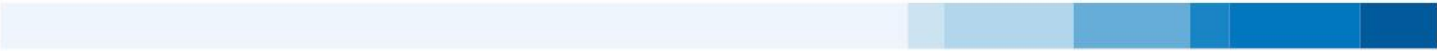
The Academy of Finland launched the Finnish Flagship Programme in 2017 to support high-quality research and stimulate the impact of research in society. This report sets a goal to analyse the development of research topics related to the selected Flagships. The development of topics was scrutinised by analysing publications from the Web of Science (WoS) Core Collection in 2005–2008 and 2015–2018. Publications reported by Flagships were used as training data for the Supervised Machine Learning model – Random Forest Binary classifier. Overall, 2,155 publications (80% for training models and 20% for testing) helped to identify 7,579 publications related to studied topics in 2005–2008 and 11,381 in 2015–2018 from the WoS database.

The bibliometric analysis studied collaboration among Finnish research organisations and international partners and measured the scientific impact of the research. Bibliometric results support the assumption that high-impact research is produced in the Flagship topics. The top 10 index, a scientific indicator measuring the share of highly-cited publications within each publication set for each Flagship-related topic, is higher than the world average for both time periods. Moreover, topics actively flourished by establishing intense national and international collaborations in their publication activity. For instance, the number of co-publications between Finnish organisations and especially with foreign partners notably increased during the period observed.

1. Introduction

[The Finnish Flagship Programme](#) was launched by the Academy of Finland in 2017. Goals of the Flagship Programme are to support high-quality research and increase the economic and societal impact emerging from the research. The Finnish Flagships aim to represent an effective mix of close cooperation with business and society, adaptability and a strong commitment from host organisations, such as universities and government research institutes.

In the first two funding calls in 2017 and 2018, six Flagships were selected: 6G Flagship – 6G Enabled Wireless Smart Society & Ecosystem, FCAI - Finnish Center for Artificial Intelligence, FinnCERES - Competence Centre for the Materials Bioeconomy, iCAN - Digital Precision Cancer Medicine Flagship, INVEST - Inequalities, Interventions and New Welfare State, and PREIN - Flagship on Photonics Research and Innovation. In the third call in 2020, the programme was supplemented by four other Flagships.



The purpose of this report is to analyse the emergence and development of research topics related to the first six Flagships before the initiation of the programme. Only topics related to the first six Flagships were included in the analysis. Those Flagships had already submitted their interim reports and training data for a machine learning model was available from the publication lists of the interim reports. Since selected topics are comprehensive and substantial, we presume that there was evolving and active research in the topics even in the years preceding the Flagship Programme.

Moreover, we are interested in what types of collaboration were in place among Flagships-related topics, which scientific impact they produced, and how these phenomena have varied in time. As a simple bibliometric indicator for scientific activity, we have used the number of publications, and the citation based impact has been measured by top 10 index.

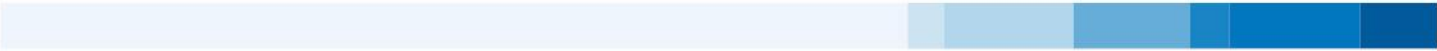
This project uses Random Forest (RF) Classifier as a machine learning technique to predict the Flagship class for the publications. Compared to the previous analysis of the Research, Development, and Innovation ecosystems in Finland (2021), which applied an Unsupervised Latent Dirichlet Allocation (LDA) model, this analysis benefits from having a training dataset with labeled Flagships which are used to run the model and see how well it performs based on accuracy indicators.

2. Data and methods

2.1. Data description

The Web of Science (WoS) Core collection administered by Clarivate Analytics was used to identify Flagships-related research papers. To identify Flagships topics from the bulk of WoS papers, this project relied on labeled data – a set of publications with known Flagship class. These papers were reported by Flagships themselves and such information as titles, keywords, and abstracts is used to train a machine learning model. The labeled dataset is split between training and test data and the accuracy of the model is assessed on the unseen by the model test data.

Specifically, in the data pre-processing step, reported papers were joined with WoS data for 2019–2021 by their unique item identifier: WoS ID or Digital Object Identifier (DOI) number. Moreover, this dataset was supplemented with six blank publications which have only keywords specific for each of the Flagships. These keywords were generated by the machine learning algorithm in the Academy of Finland, using the text of funding applications of the first six Flagships as data. Machine learning algorithm produced preliminary lists of keywords which were then cleaned and supplemented by the Academy of Finland's science advisers familiar with research topics of the six Flagships. See Appendix A for lists of final keywords.



Overall, the dataset with known Flagship classes is composed of 2,155 publications. The data is not balanced between six classes because we had different numbers of papers reported by Flagships and it was not possible to find WoS or DOI identification for all of them. Identification, however, was required to join with WoS data and retrieve complete keywords and abstracts for each publication.

The textual data required pre-processing steps. First, for each publication the title, keywords, and abstract were joined together into publication information column. Second, we removed punctuation, lowered all words, removed numbers and stop words (the most common words in English), and stemmed the data (natural language processing method to lower inflection in words to their roots).

Next, we created a corpus of documents, a special object storing the collection of texts. The corpus allows us to tokenize words on unigrams and bigrams (one and two-words sequences) as well as to create a document term matrix. Unigrams were combined with bigrams to overcome the bag-of-words limitation which assumes that the order of words does not matter. Thus, accounting for pairs of words - bigrams, helps to consider meaningful relations between words (Wallach, 2006).

A document term matrix is another form of textual representation showing statistics of the word's importance in the corpus (Silge & Robinson, 2017). TF-IDF (term frequency and inverse document frequency) raises the importance of more specific words of collection and decreases the weight of commonly used words.

Overall, these preprocessing steps were first applied to the 2,155 publications, of which 80% was used for training models and 20% for testing their accuracy. Moreover, the TF-IDF matrix containing words from the labeled dataset was later used for the new creating matrix from WoS data as the dictionary.

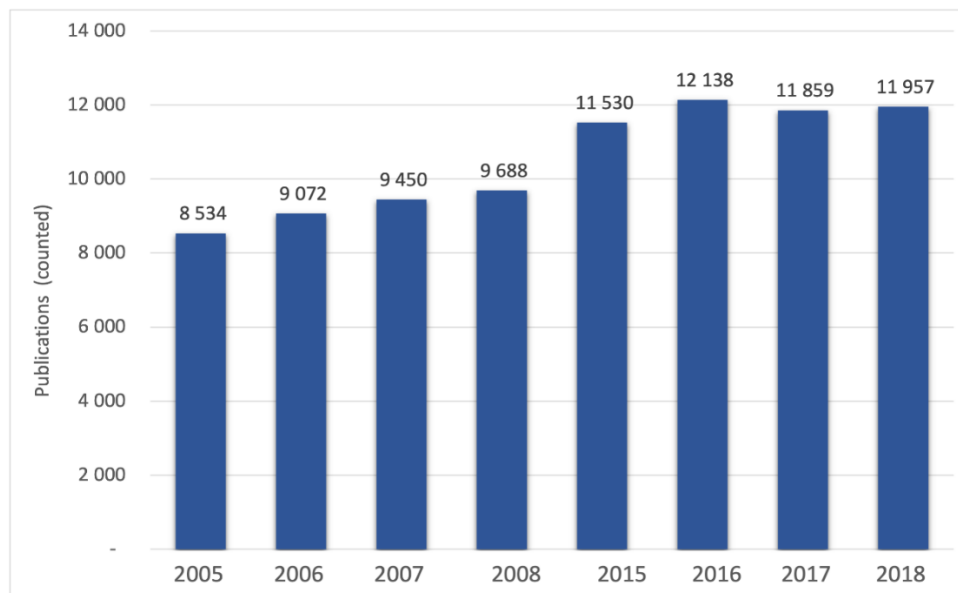
To analyze the evolution of Flagships' topics, WoS data from the years 2005–2008 and 2015–2018 was used. These time spans were chosen to compare the development of the topics in relatively long four-year periods. Years 2005–2008 represent an earlier stage and years 2015–2018 a stage before the launch of the programme.

Overall, there were 84,228 publications which met the following criteria: their document type was article, proceedings paper, meeting abstract, review or letter and their publication country was Finland. Since the purpose of the project is to analyze the development of research topics related to the Flagship Programme funded by the Academy of Finland, we filtered dataset for Finnish publications. A publication was considered Finnish if at least one author was associated with a Finnish organization. Moreover, the dataset contains only English language publications.

Table 1. Publication types in the dataset.

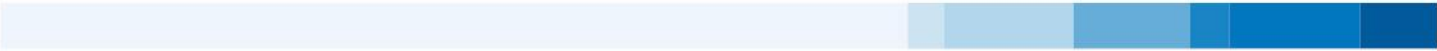
| Document type | Count |
|-------------------|--------|
| Article | 59,391 |
| Proceedings Paper | 13,527 |
| Meeting Abstract | 7,717 |
| Review | 2,605 |
| Letter | 988 |

The natural language preprocessing techniques were also applied to this bulk of papers. In the last step of creating the document term matrix, we used the dictionary of words of the training data for filtering. This allowed us to apply the same trained model to the new data.

**Figure 1. Number of publications in the dataset by year.**

2.2. Modeling

This project uses the Random Forest Binary Classifiers as a method. The choice of the method in detail and overview of related works is explained in the Appendix B. Random forests proved to be a well-performed classification method and also rather simple in results interpretation (Boehmke & Greenwell, 2019, p. 203). Being a compound of an ensemble of trees, Random Forests are an extension of the bagging technique. Bagging presumes averaging predictions of individual learners by reducing their variance (p. 192). While this aggregation process can be applied to any type of method, it proves to be especially effective for high variance and unstable learners such as KNN and Decision Trees (p. 192). However, bagging trees also impact their



correlation and weaken the effects of variance reduction (p. 204). Random Forests tend to overcome this problem by performing split-variable randomization at any split point (p. 204). Thus, Random Forests represent the de-correlated option of bagged decision trees.

Binary Random Trees Classifier predicts a probability with which a document belongs to a class. Thus, by setting a probability threshold above 0.5, the model's results become easily understandable. Binary models predicting if a publication belongs to a class or not were built separately for every Flagship topic. A document can belong to several Flagship topics simultaneously. For example, FCAI and 6G topics have common terms and can have overlapping publications.

Modeling had the following workflow. After preprocessing the labeled dataset of 2,155 publications, we assigned 80% of the data as training data and 20% as test data for model's performance. Before running every separate Flagship model, we recoded the targeted Flagship topic as 'yes' and all others as 'no'. Thus, every model had 1,724 training observations. Based on the 20% of test data where a class is known, we can assess the performance of the binary classifiers by looking at the accuracy score (see Appendix A). The accuracy score is a metric showing the number of correct predictions divided by the total number of predictions (Boehmke & Greenwell, 2019).

Random Forest has the following hyperparameters that have a higher impact on the model's performance that could be tuned: the number of trees in the forest to grow, the number of features to take into account at any split - mtry, cut off scheme (Boehmke & Greenwell, 2019). The models were automatically fitted with 500 trees each, mtry was set to 39, cut off set to 0.5/0.5 meaning that the winning class for observation was made on this probability. The search for other hyperparameters was tried though the expand grid but it did not lead to better results than the default models. Boehmke and Greenwell point out that default hyperparameters values generally tend to produce good results and the method requires little application of tuning.

Before proceeding with the results, this project attempted to apply a Multi-class Random Forest classifier. It means that one model was fitted for all Flagships. While the achieved accuracy was also high at 0.81, it was lower than most of the individual classifiers. Besides, interpreting the model's results was less straightforward.

Overall, the following number of publications from the selected years in the WoS database were identified per each Flagship topic (see Figure 2). These results were selectively checked and approved by experts in the Academy of Finland.

Specifically, experts from every Flagship's topic went through a subset of identified publications to analyze if the publication titles were relevant to the topic. After the expert checking only 127 works from all topics were considered non-relevant and were excluded from the final dataset.

Results were also checked by comparing the number of identified publications per Flagship topic with the number of publications per scientific field. While topics cannot be directly mapped to a science area, in this project we compared iCAN topics or cancer medicine with the development of Medicine and Health science area which are relatively close to each other. The reason for comparison is based on the highest number of iCAN-related publications identified by the model. To sum up, we considered the model's result relevant for iCAN topics since, according to [Vipunen](#) statistical service, there were on average around 3,700 Medicine and Health publications per year in 2005–2008 and 4,800 in 2015–2018.

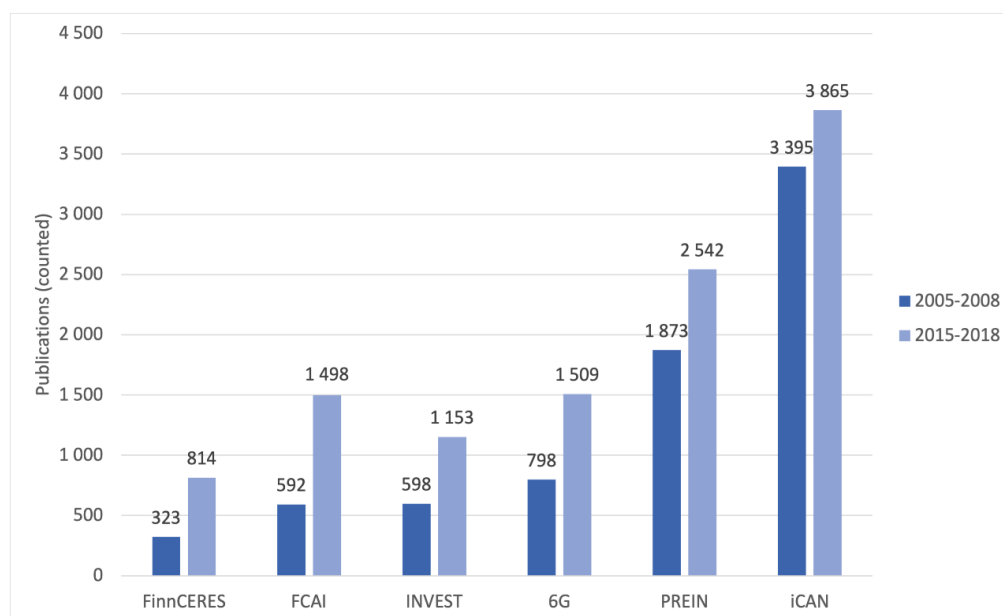


Figure 2. Number of publications identified by the model per Flagship topic and period.

3. Bibliometric analysis

In order to evaluate the development of scientific activity in Finland within the larger subject fields based on the Flagship topics, we shall consider several statistical indicators. As explained in the previous chapter, our data consist not only of the publications from the organizations actively participating in the first six Flagships but of all Finnish publications connected to the Flagship themes. All analyses were carried out for two 4-year periods. The latter period between 2015–2018 corresponds with the years just before the launch of the Flagship Programme, whereas the earlier period from 2005 to 2008 is used for comparison. Naturally, a future follow-up study covering years 2020–2023, for instance, would give valuable information on the actual effects of the Flagship Programme.

3.1. Number of publications

The overall level of research activity can be estimated using publication volumes, shown in Table 2. We have removed a small number of publications that were not suitable for bibliometric analysis from the raw data produced by machine learning method. Therefore, publication numbers in Table 2 differ slightly from those given in Figure 2. We also note that the number of publications on each topic is sufficient for a meaningful analysis of scientific impact.

The amount of publications in years 2005–2008 show that there has already been active research connected to all Flagship topics a decade before the launch of the Flagship Programme. Relative increase has been most pronounced in subject fields close to topics represented by FCAI and FinnCERES. INVEST and 6G topics have had an almost twofold growth in publication volumes. For research connected to PREIN and iCAN that already had considerable numbers of publications in 2005–2008, the growth rate has been clearly lower.

Table 2. Number of publications for each Flagship topic.

| Flagship topic | 2005–2008 | 2015–2018 | Change | Relative growth |
|----------------|-----------|-----------|--------|-----------------|
| FCAI | 590 | 1,497 | 907 | 153.7% |
| INVEST | 552 | 1,099 | 547 | 99.1% |
| PREIN | 1,871 | 2,542 | 671 | 35.9% |
| FinnCERES | 322 | 812 | 490 | 152.2% |
| iCAN | 3,393 | 3,864 | 471 | 13.9% |
| 6G | 797 | 1,509 | 712 | 89.3% |

Part of the apparent growth is due to the overall expansion of the Web of Science database. Most Flagship topics cover only a small portion of the total volume of all Finnish publications, as shown in Table 3. The topics with the most publications, PREIN and iCAN, actually represent lower shares of the total volume of all Finnish publications during the years 2015–2018 than in 2005–2008.

Table 3. Relative share of Flagship topic publications among all Finnish publications.

| Flagship topic | 2005–2008 | 2015–2018 |
|----------------|-----------|-----------|
| FCAI | 1.28% | 1.96% |
| INVEST | 1.20% | 1.44% |
| PREIN | 4.06% | 3.32% |
| FinnCERES | 0.70% | 1.06% |
| iCAN | 7.36% | 5.05% |
| 6G | 1.73% | 1.97% |

3.2. Collaboration

One of the goals of the Flagship Programme has been to increase the societal impact of research through direct interaction between the academic community and society. On the other hand, the programme emphasizes the importance of international collaboration. In Table 4, we show how the publications on Flagship topics are divided into three separate categories representing different forms of scientific collaboration: 1) international collaboration, with researchers from both Finland and abroad, 2) domestic collaboration between two or more Finnish organizations, and 3) publications with authors from just one organization. The results for all Finnish publications under respective periods are given for comparison on the bottom row.

Table 4. Distribution of publications according to type of collaboration.

| Flagship topic | International 2005–2008 | National 2005–2008 | Single organization 2005–2008 | International 2015–2018 | National 2015–2018 | Single organization 2015–2018 |
|----------------------------|----------------------------|-----------------------|-------------------------------------|----------------------------|-----------------------|-------------------------------------|
| FCAI | 19% | 27% | 54% | 33% | 25% | 42% |
| INVEST | 21% | 48% | 31% | 30% | 44% | 26% |
| PREIN | 35% | 20% | 46% | 52% | 17% | 31% |
| FinnCERES | 23% | 27% | 51% | 40% | 24% | 36% |
| iCAN | 29% | 49% | 21% | 37% | 53% | 11% |
| 6G | 15% | 17% | 68% | 40% | 17% | 43% |
| Finland average | 43% | 22% | 36% | 61% | 17% | 22% |

Even though the share of internationally co-published papers has clearly increased in all Flagship topics during the span of observation, the shares

remain below the overall national average for both time periods. Again, we observe that the Flagship topics could be classified into three groups. For FCAI and 6G topics, publications by a single organization are most common, whereas for topics related to INVEST and iCAN, collaboration between Finnish organizations (typically involving hospitals) continues to be the dominant form of collaboration. Within PREIN and FinnCERES topics, there has been a notable shift from largely single organization publications to strong international collaboration.

We have also examined the share of publications with at least one Finnish industrial or commercial partner (see Table 5 below). As expected, industrial collaboration is more common in the technology-oriented Flagship topics FinnCERES and 6G. For iCAN topic, the share of industrial collaboration has increased considerably between the two time periods. Part of this is explained by the increased outsourcing of some hospital activities, such as laboratory services and medical imaging.

Table 5. Collaboration with domestic industrial partners.

| Flagship topic | 2005–2008 | 2015–2018 |
|------------------------|-------------|-------------|
| FCAI | 5.9% | 7.2% |
| INVEST | 2.2% | 4.1% |
| PREIN | 6.1% | 6.2% |
| FinnCERES | 13.0% | 9.5% |
| iCAN | 6.0% | 11.2% |
| 6G | 13.8% | 13.1% |
| Finland average | 6.2% | 6.4% |

3.3. Subject fields

Publications in the Web of Science database are labelled with one or more tags indicating their subject fields. In the following tables we have collected for each Flagship topic the 10 most frequently appearing Web of Science subject fields and their shares of the total publication volume in the Flagship topic. The results follow closely the proposed Flagship research areas.

Table 6. Most prevalent Web of Science subject fields in FCAI topic.

| FCAI | 2005–2008 | 2015–2018 |
|--|-----------|-----------|
| Computer Science, Artificial Intelligence | 14.5% | 9.4% |
| Engineering, Electrical & Electronic | 3.2% | 7.4% |
| Computer Science, Theory & Methods | 5.5% | 5.9% |
| Education & Educational Research | | 5.8% |
| Computer Science, Information Systems | 5.1% | 5.7% |
| Computer Science, Software Engineering | 6.4% | 4.9% |
| Management | | 2.5% |
| Computer Science, Interdisciplinary Applications | 3.9% | 2.5% |
| Statistics & Probability | 2.2% | 2.4% |
| Public, Environmental & Occupational Health | | 1.7% |
| Ecology | 2.7% | |
| Mathematics, Applied | 2.1% | |
| Mathematical & Computational Biology | 1.9% | |

Flagship FCAI was set up to develop artificial intelligence and its applications to real-world problems. In addition to many subfields of computer science, the keywords typical for the FCAI topic appear in some health and biology related publications.

Table 7. Most prevalent Web of Science subject fields in INVEST topic.

| INVEST | 2005–2008 | 2015–2018 |
|---|-----------|-----------|
| Public, Environmental & Occupational Health | 13.5% | 13.8% |
| Psychiatry | 12.9% | 9.7% |
| Pediatrics | 7.2% | 5.5% |
| Psychology, Developmental | 4.5% | 5.3% |
| Education & Educational Research | 2.2% | 4.5% |
| Obstetrics & Gynecology | 4.0% | 3.5% |
| Sociology | | 2.3% |
| Nursing | 3.3% | 2.2% |
| Multidisciplinary Sciences | | 2.1% |
| Psychology, Educational | | 2.0% |
| Endocrinology & Metabolism | 3.7% | |
| Psychology, Multidisciplinary | 3.4% | |
| Clinical Neurology | 2.2% | |

Flagship INVEST seeks means to improve the wellbeing of younger generations and develop new ideas for the welfare state. The observed subject fields correspond well with these themes.

Table 8. Most prevalent Web of Science subject fields in PREIN topic.

| PREIN | 2005–2008 | 2015–2018 |
|---------------------------------------|-----------|-----------|
| Optics | 17.6% | 12.9% |
| Physics, Applied | 11.5% | 10.9% |
| Materials Science, Multidisciplinary | 5.5% | 8.3% |
| Engineering, Electrical & Electronic | 6.1% | 6.1% |
| Chemistry, Multidisciplinary | | 4.8% |
| Chemistry, Physical | 5.2% | 4.7% |
| Physics, Condensed Matter | 5.5% | 3.7% |
| Astronomy & Astrophysics | 2.9% | 3.7% |
| Nanoscience & Nanotechnology | | 3.4% |
| Multidisciplinary Sciences | | 3.2% |
| Physics, Multidisciplinary | 3.4% | |
| Physics, Atomic, Molecular & Chemical | 2.8% | |
| Chemistry, Analytical | 2.3% | |

The object of PREIN is to promote research collaboration between diverse partners in the field of photonics, i.e., light-based technologies. In addition to optics, the publications represent various subfields in physics and chemistry.

Table 9. Most prevalent Web of Science subject fields in FinnCERES topic.

| FinnCERES | 2005–2008 | 2015–2018 |
|--------------------------------------|-----------|-----------|
| Materials Science, Paper & Wood | 37.3% | 17.1% |
| Chemistry, Multidisciplinary | 4.9% | 14.4% |
| Polymer Science | 4.2% | 10.6% |
| Engineering, Chemical | 3.9% | 6.8% |
| Materials Science, Multidisciplinary | | 3.8% |
| Materials Science, Textiles | | 3.6% |
| Chemistry, Physical | 5.2% | 3.5% |
| Energy & Fuels | 2.2% | 3.4% |
| Forestry | 8.2% | 3.0% |
| Biotechnology & Applied Microbiology | 3.6% | 3.0% |
| Environmental Sciences | 4.0% | |
| Biochemistry & Molecular Biology | 2.0% | |

FinnCERES concentrates on developing renewable materials for a sustainable economy. The Web of Science subject fields suggest the application areas come from cellulose-based industries.

Table 10. Most prevalent Web of Science subject fields in iCAN topic.

| iCAN | 2005–2008 | 2015–2018 |
|---|-----------|-----------|
| Oncology | 19.5% | 16.8% |
| Surgery | 5.0% | 6.2% |
| Multidisciplinary Sciences | | 5.0% |
| Cardiac & Cardiovascular System | 2.7% | 3.7% |
| Cell Biology | 3.2% | 3.6% |
| Clinical Neurology | | 3.1% |
| Gastroenterology & Hepatology | 3.9% | 2.9% |
| Biochemistry & Molecular Biology | 3.6% | 2.9% |
| Immunology | 3.5% | 2.8% |
| Radiology, Nuclear Medicine & Medical Imaging | | 2.4% |
| Endocrinology & Metabolism | 4.1% | |
| Pathology | 2.9% | |
| Genetics & Heredity | 2.7% | |

The mission of the iCAN Flagship is to find new pivotal treatments for certain specific types of cancer, one of the fields with the most publications in Web of Science.

Table 11. Most prevalent Web of Science subject fields in 6G topic.

| 6G | 2005–2008 | 2015–2018 |
|--|-----------|-----------|
| Engineering, Electrical & Electronic | 32.8% | 36.8% |
| Telecommunications | 26.2% | 23.9% |
| Computer Science, Information Systems | 4.7% | 7.1% |
| Computer Science, Artificial Intelligence | 4.0% | 4.9% |
| Computer Science, Theory & Methods | 5.0% | 4.0% |
| Computer Science, Hardware & Architecture | 5.2% | 3.5% |
| Transportation Science & Technology | 2.4% | 2.2% |
| Computer Science, Software Engineering | 1.8% | 1.6% |
| Computer Science, Interdisciplinary Applications | 1.6% | 1.6% |
| Imaging Science & Photographic Technology | 3.1% | 1.4% |

6G is clearly the most specialized Flagship topic, if measured by the share of the two most frequent subject fields. Around 60% of all publications represent either electrical engineering or telecommunications.

3.4. Scientific impact

The scientific impact of the topics was analysed by the top 10 index. This indicator measures the share of highly-cited publications within a publication set under scrutiny. More precisely, it expresses the ratio of the share of publications that belong to the most cited 10% in their respective fields and year of publication to the expected share of 10%. Thus, a top 10 index value of 1.0 signifies an impact on a par with the world average, and values above 1.0 mark research with higher than average impact. The top 10 index results for each Flagship topic are presented in Table 12.

All Flagship topics exceed the world average during both time periods, and almost all topics also have a higher impact than Finnish contemporaneous publications in general. Publications related to artificial intelligence (FCAI) have had a stable high level impact, whereas two other technology-oriented topics (FinnCERES and 6G) have experienced a notable increase in their impact between the two time periods. Top 10 indices for the rest of Flagship topics have remained rather stable between 2005-2008 and 2015-2018.

Table 12. Top 10 index for Flagship topics.

| Flagship topic | 2005–2008 | 2015–2018 |
|------------------------|-------------|-------------|
| FCAI | 1.40 | 1.45 |
| INVEST | 1.10 | 1.11 |
| PREIN | 1.12 | 1.11 |
| FinnCERES | 1.30 | 1.72 |
| iCAN | 1.13 | 1.21 |
| 6G | 1.09 | 1.71 |
| Finland average | 1.04 | 1.13 |

3.5. Collaboration networks

In order to recognize the most productive organizations and to illuminate their collaboration networks, we provide two graphs for every Flagship topic, one for each time period (Figures 3–8). The relevant organizations are depicted as circles. The area of a circle is proportional to the publication volume of the organization (each graph is scaled separately). The colour of the circle depends on the top 10 index of the organization's publications.

The arcs connecting the circles denote co-publications between different organizations, with thickness of the arc being proportional to the number of co-publications and the colour signalling the top 10 index. A circle or arc left grey means that the underlying publication set, even though scientific, does not contain enough publications suitable for impact analysis.

In addition to the named organizations, there are some collective entities. **Other HEI** refers to other institutes of higher education, typically with a very specific research area, e.g., Helsinki Institute of Information Technology. **Industry** represents all industrial or commercial Finnish partners, **Other** refers to all non-commercial Finnish organizations that are not present in the graph under a named entry, and **Foreign** contains any collaborating partners from abroad. The total number of publications and their overall top 10 index is given on the left upper corner of each figure.

In all research themes constructed around the Flagship topics, the host organizations of first six Flagships are among the most visible actors during the period 2015–2018. The absolute number of co-publications between Finnish organizations and especially the number of co-publications with foreign partners has grown during the observation period. Therefore the arcs connecting the organizations are usually wider in the graphs on the right representing the years 2015–2018. Another recurring feature is the significance of international collaboration for the scientific impact, i.e., the circle presenting foreign organizations tends to glow in a deeper hue of blue.

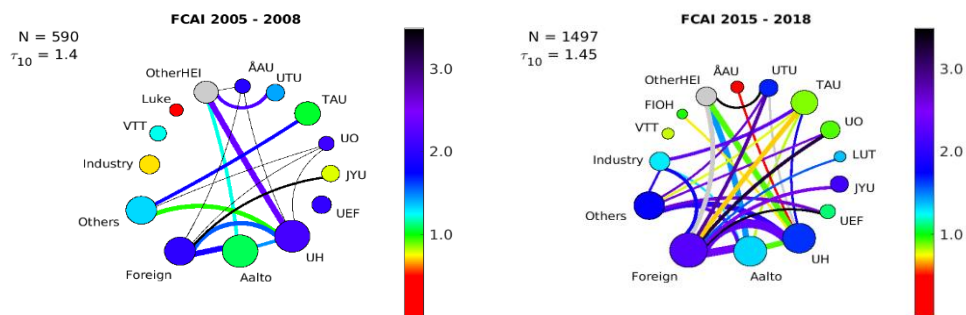


Figure 3. Collaboration networks in FCAI topic.

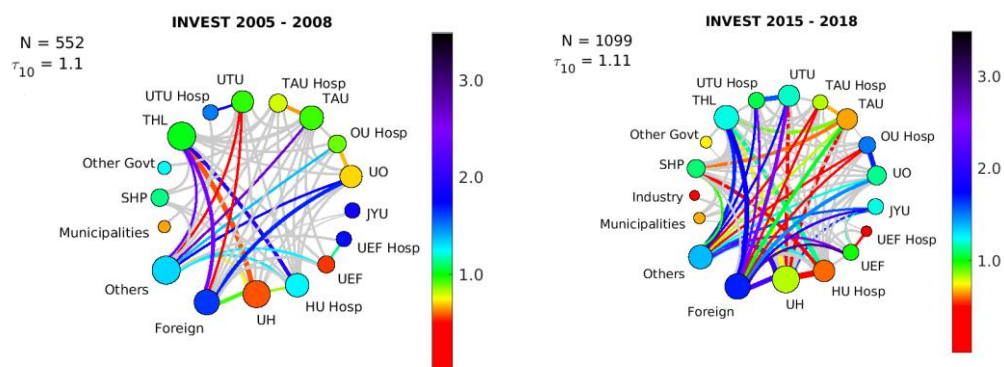


Figure 4. Collaboration networks in INVEST topic.

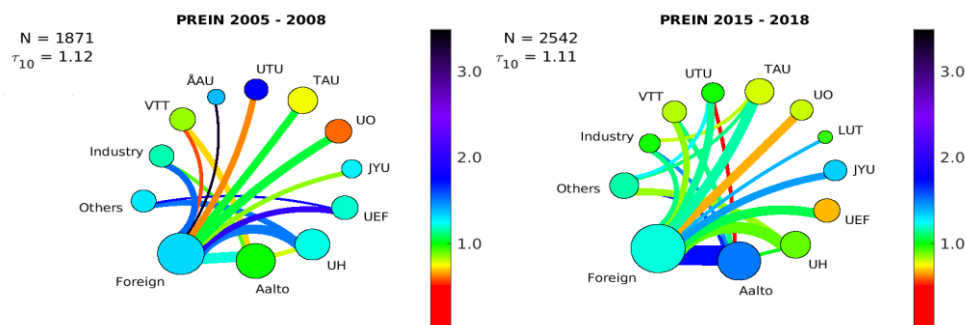


Figure 5. Collaboration networks in PREIN topic.

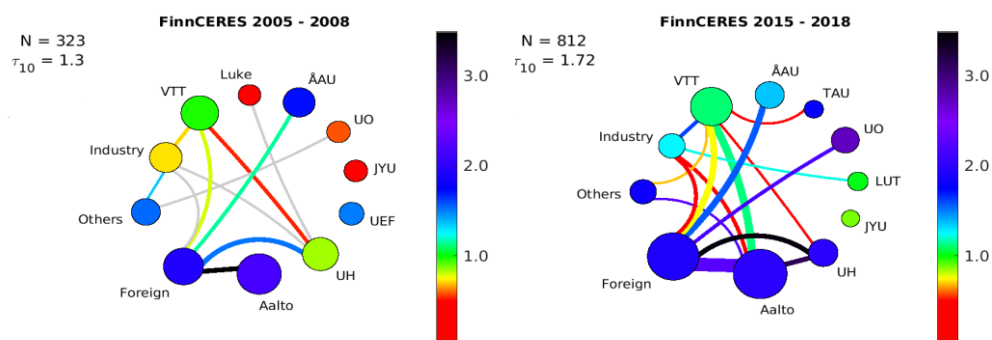


Figure 6. Collaboration networks in FinnCERES topic.

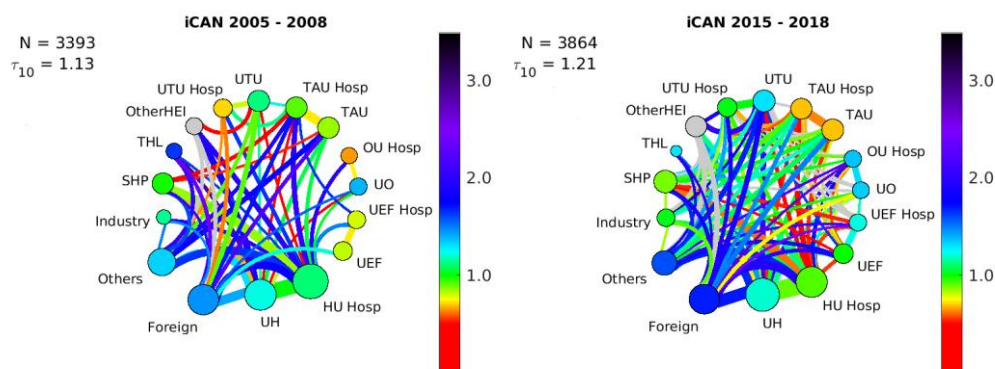


Figure 7. Collaboration networks in iCAN topic.

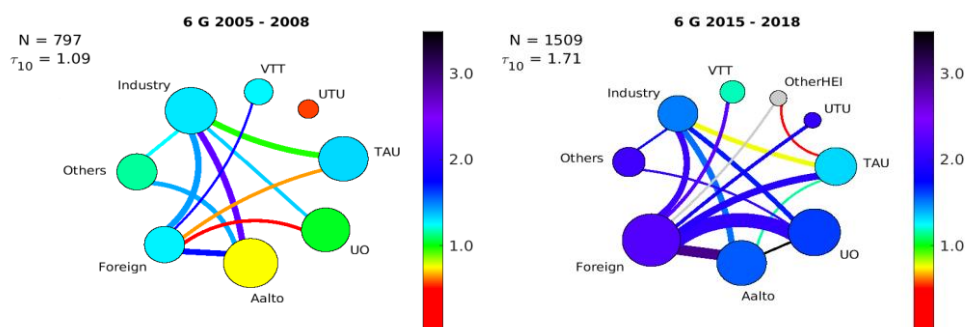


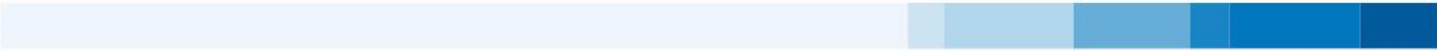
Figure 8. Collaboration networks in 6G topic.

4. Conclusions

This project aimed to analyze the development of Flagship-related topics before the initiation of the Flagship Programme by the Academy of Finland in 2017. To find the publications connected to the Flagship topics, the Random Forest Binary Classifier machine learning technique was applied to a set of predefined 2,155 publications where Flagship class is known and labeled. This set of publications was provided by Flagships themselves via reports, data was pre-processed following natural language techniques for textual data, and titles, keywords, and abstracts from publications were used to train the models.

Next, bibliometric analysis was performed for publications from the Web of Science collection that were identified by the models. Collaboration networks and subject fields per Flagship topic were detected and the scientific impact of the topics was calculated using the top 10 index.

Overall, the results show an increase in international collaboration for all topics in the observed time periods. Publications associated with each Flagship topic had different amounts of national and international collaboration. The largest share of publications associated with FCAI and 6G topics were publications by a single organization. Publications associated with INVEST and iCAN topics had the largest share of national collaboration between Finnish



organizations. Publications associated with PREIN and FinnCERES topics had intensified international collaboration in 2015–2018 compared to 2005–2008.

Notably, the scientific impact of all Flagships-related publications exceeded the world average in both periods, which supports the hypothesis that topics chosen for the Finnish Flagship Programme are significant and were actively developing in Finland even before the launch of the programme.

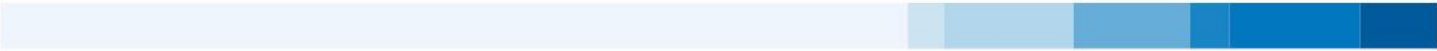
Despite the high accuracy of the Random Forest models and checking the results by experts, the analysis has some limitations. First of all, the training data for the models was unbalanced with the number of publications per Flagship from their interim reports. For instance, INVEST had the lowest number of publications, which might influence how many papers were found by the model for this topic since the algorithm had a narrower vocabulary of words to learn from. Secondly, the terms belonging to each topic have an impact on the results, since some Flagship topics as FCAI and 6G have more overlapping terms with each other than others. This creates ambiguity for the model to identify a Flagship class.

Another potential weakness lies in the use of Web of Science database. Although excellent in many fields of natural sciences, the coverage of Web of Science is less optimal for analyses in the social sciences or humanities, especially when it comes to publications in other languages than English. This may have affected the bibliometric results in INVEST Flagship topic.

Overall, this analysis presents an appealing direction for future research. For example, a similar analysis could be conducted in a few years time to capture the impact of the Flagship Programme.

References

- Academy of Finland, *Finnish Flagship Programme*. [Online] Available at: <https://www.aka.fi/en/research-funding/programmes-and-other-funding-schemes/flagship-programme/>
- Mankinen, K., & Leino, Y., 2021. *Identifying research topics and collaboration networks in Finland: topic modelling of scientific publications in 2008–2019*. [Online] Available at: <https://www.aka.fi/globalassets/2-suomen-akatemia-toiminta/4-julkaisut/julkaisut/identifying-research-topics-and-collaboration-networks-in-finland.pdf>
- Pranckevičius, T., & Marcinkevičius, V., 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2), pp. 221–232.
- Wang, Q., Luo, Z., Huang, J., Feng, Y., & Liu, Z., 2017. A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM. *Computational intelligence and neuroscience*.
- Bonaccorso, G., 2018. *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.
- Kandimalla, B., Rohatgi, S., Wu, J., & Giles, C. L., 2021. Large scale subject category classification of scholarly papers with deep attentive neural networks. *Frontiers in research metrics and analytics*, 5, 600382.
- Beltagy, I., Lo, K., & Cohan, A., 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint:1903.10676*.
- Robinson, D., 2022. *LDA Topic Model Instability*. [Online] Available at: <https://towardsdatascience.com/lda-topic-model-instability-c2fedb77d249>
- Diaz-Valenzuela, I., Martin-Bautista, M. J., & Vila, M. A., 2014. A fuzzy semi-supervised clustering method: application to the classification of scientific publications. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, Cham, pp. 179–188.
- Eshima, S., Imai, K., & Sasaki, T., 2020. Keyword assisted topic models. *arXiv preprint:2004.05964*.
- Klassen, M., & Paturi, N., 2010. Web document classification by keywords using random forests. In *International Conference on Networked Digital Technologies*. Springer, Berlin, Heidelberg, pp. 256–261.
- Xu, B., Guo, X., Ye, Y., & Cheng, J., 2012. An improved random forest classifier for text categorization. *J. Comput.*, 7(12), pp. 2913–2920.



Wallach, H. M., 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pp. 977–984.

Silge, J., & Robinson, D., 2017. *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

Boehmke, B., & Greenwell, B., 2019. *Hands-on machine learning with R*. Chapman and Hall/CRC.

Vipunen, *Bibliometriikka (Web of Science)*. [Online] Available at: <https://vipunen.fi/fi-fi/kkyhteiset/Sivut/Bibliometriikka.aspx>

Appendix A. Keywords and publications by Flagship topic

| Flagship | Keywords | Publications in training da- taset | Binary model's accuracy | Identified publications |
|-----------|--|--|-------------------------------|----------------------------|
| FCAI | bayesian, trust, privacy, ethical ai, machine learning, computational data analysis, artificial intelligence, probabilistic modelling, deep learning, security and privacy, interactive machine learning, autonomous, big data, computer vision, data mining, decision making, explainable ai, facial recognition, inference, natural language processing, neural network, pattern recognition, personalized medicine, personalised medicine, predictive model, robotics, un-supervised learning, unsupervised learning, supervised learning, reinforcement learning | 569 | 0.85 | 2,087 |
| FinnCERES | biofuel, biocomposite, sustainability, sustainable, biowaste, bio-material, biorefin, wearable, packaging, lignin, nanocellulose, cellulosic, biopolymer, lignocellulos, functional surface, biomass, bioplastic, renewable, fiber, fibre, forest, sustainable production, bioeconomy, cellulose, cellulos material, renewable material, bio-based material, bio-based, pulping, hemicellulose, textile | 221 | 0.98 | 1,135 |
| iCAN | precision medicine, biomarker, drug screening, cancer, organoids, ovarian cancer, patient empowerment, colorectal, leukemia, big data, health data, biobank, tumour microenvironment, digital health, diagnostic, breast cancer, clinical trial, tumor, health technology, personalized medicine, personalised medicine | 412 | 0.97 | 7,257 |
| INVEST | welfare state, social institution, life course, public health, social environment, famil, skill development, wellbeing, child, youth, trajector, transition, social service, health service, social protection, mental health problem, health problem, disruptive behavior, social status, depression, teacher, social inequality, social equality, bullying, employment, school, social security, intervention, income inequality, income equality, inequality, equality, social inclusion, gene family | 165 | 0.95 | 1,682 |
| PREIN | photonics, optics, optoelectronics, nanophotonics, nonlinear optics, quantum optics, quantum photonics, silicon photonics, ultrafast optics, ultrafast dynamics, laser, lasers, plasmon, plasmonics, spectroscopy, imaging, non-classical light, waveguide, fiber optics, optical fiber, nanowire, frequency comb, solar cell, single-photon source, laser diode, photodetector, metamaterials | 466 | 0.92 | 4,413 |
| sixG | wireless, data analytics, antenna, iot, internet of things, telecommunications, transceivers, transmission, digitalisation, operator, thz, optimisation, ubiquitous, latency, verticals, spectrum, algorithm, cloud computing, server, connectivity, transceiver, sensor, ubiquitous computing, electronics, radio, 5g, 6g, ghz, mmwave, antennas, edge computing | 322 | 0.92 | 2,306 |

Appendix B. Literature review of machine learning methods for textual data

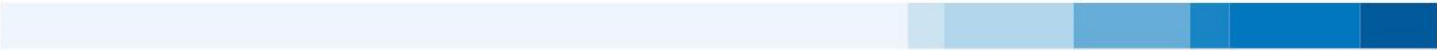
Textual data analysis can be performed by applying different machine learning methods. For instance, documents can be categorized using Support Vector Machines (SVM), Decision Trees, Logistic Regression, and Naïve Bayes when the label of the class is known in the training dataset. It is debatable which model works best. A bulk of research has applied SVMs indicating them to be one of the best classifiers (Pranckevičius & Marcinkevičius, 2017, p. 224).

While SVMs are adaptive and robust, they are also quite complex (Luo, et al., 2017, p. 290). Depending on the analysis and available data, other methods such as Logistic Regression proved to give more accurate predictions than SVM, Naïve Bayes, or Decision Trees (Pranckevičius & Marcinkevičius, 2017).

With the rising popularity of Deep Learning methods, such techniques as Deep Neural Networks have been used to classify textual data, too. However, this approach requires much more data and bigger algorithm capacities. For instance, an application of the Deep Neural Network to the WoS database used 45 million observations where the model outperformed SVM, Logistic Regression, Random Forest and Naïve Bayes applied to the same dataset (Kandimalla, et al., 2021). Training such a model also requires a large amount of labeled data (Beltagy, et al., 2019). In this regard, it's worth comparing methods that have a closer algorithm capacity (Pranckevičius & Marcinkevičius, 2017, p. 222).

Since research often deals with unlabeled textual data, Unsupervised Learning methods such as Clustering and Topic Modeling became common tools to be applied. Among Topic Modeling, Latent Dirichlet Allocation (LDA), Latent Semantic Analysis, and Probabilistic Latent Semantic Analysis are commonly used (Bonaccorso, 2018). Despite the popularity of the LDA algorithm, one should be aware of its instability and absence of a common truth on how to evaluate the model's performance and choose the number of topics (Robinson, 2022).

One way to overcome this shortcoming is to use expert knowledge to create a topic vocabulary and evaluate distinguished topics, but this can be expensive and time-consuming. Some clustering variations, such as semi-supervised clustering, which use a small amount of information for the clustering process, were proposed by researchers (Diaz-Valenzuela, et al., 2014). For instance, expert's knowledge is used to determine instance-level constraints for optimum cluster partition. An expert with detailed knowledge of the input data provides these constraints, which help to find an optimum cut for a dendrogram (Diaz-Valenzuela, et al., 2014).



Closer assistance is also presumed in keyword-assisted topic modeling. As such, providing a short list of keywords before fitting a topic model proves to produce more meaningful results than LDA (Eshima, et al., 2020).

This analysis benefits from having a training dataset with known Flagship classes. Thus, we limit our scope to Supervised Machine Learning methods and do not consider Deep Neural Networks due to the relatively small amount of data available. From this bulk of methods, the Random Forest classifier has proven to advance classification accuracy and to be a convenient and effective method (Klassen & Paturi, 2010; Luo, 2017). Also, research demonstrates that Random Forest often tends to outperform SVM, KNN, and Näive Bayes (Xu, et al., 2012).

Appendix C. Abbreviations for organizations

| | |
|----------------|---|
| Aalto | Aalto University |
| UH | University of Helsinki |
| HU Hosp | Helsinki University Central Hospital |
| UEF | University of Eastern Finland |
| UEF Hosp | University of Eastern Finland Hospital |
| JYU | University of Jyväskylä |
| ULA | University of Lapland |
| LUT | Lappeenranta-Lahti University of Technology |
| DefUni | National Defence University |
| UO | University of Oulu |
| OU Hosp | Oulu University Hospital |
| Hanken | Hanken School of Economics |
| UNIARTS | University of the Arts |
| TAU | University of Tampere |
| TAU Hosp | Tampere University Hospital |
| UTU | University of Turku |
| UTU Hosp | Turku University Hospital |
| UVA | University of Vaasa |
| ÅAU | Abo Akademi University |
| OtherHEI | Other Institute of Higher Education |
| FFA | Finnish Food Authority |
| GSF | Geological Survey of Finland |
| FMI | Finnish Meteorological Institute |
| Luke | National Resources Institute Finland |
| NLS | National Land Survey of Finland |
| SYKE | Finnish Environment Institute |
| RNSA | Radiation and Nuclear Safety Authority in Finland |
| THL | Finnish Institute for Health and Welfare |
| FIOH | Finnish Institute of Occupational Health |
| VATT | VATT Institute for Economic Research |
| VTT | VTT Technical Research Centre of Finland |
| Other Govt | Other Government Institutes |
| SHP | Other hospitals |
| Industry | Industry |
| Others | Others |
| Foreign | Foreign Organizations |
| Municipalities | Municipalities |