EVALUATION ON ACCURACY OF MAPPING SCIENCE TO THE UNITED NATIONS' SUSTAINABLE DEVELOPMENT GOALS (SDGS) OF THE AURORA SDG QUERIES

Felix Schmidt
University Library
University of Duisburg-Essen
Duisburg, Germany
felix.schmidt@uni-due.de
ORCID: 0000-0002-9277-7954

Maurice Vanderfeesten
University Library
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
maurice.vanderfeesten@vu.nl
ORCID: 0000-0002-5119-3514

June 16, 2021

ABSTRACT

In this study we evaluate the accuracy of our Aurora SDG classification model version 5, to match research papers to the Sustainable Development Goals (SDGs) of the United Nations. The aim of this investigation is to be transparent about the accuracy of the model, and enable use of the model in reporting and strategy analysis by University leadership. The measurements are based on a baseline, 'golden set', where researchers of the Aurora universities handpicked publications that relate to an SDG. We measured the precision and recall of the Aurora model, and related it to our previous version of the model, and to the Elsevier SDG model.

Keywords Bibliometrics · SDGs · Sustainable Development Goals · Precision · Recall · Scientific publications · Mapping · Boolean search queries · Scopus

1 Introduction

The Aurora Universities network is a network of nine leading European research universities founded in 2016, united by their commitment to match academic excellence with creating societal impact. University leadership asked the Aurora universities to demonstrate the relevance of their research to grand societal challenges [1, 2, 3]. Therefore the Sustainable development goals (SDGs) of the United Nations were chosen as a framework to be the leading narrative, to match research papers to these topics using boolean search queries [4]. The bibliometricians of the Aurora universities have worked together since 2017 to create our own definition of the *"Aurora SDG Queries"* matching our research to the SDGs. Since the start of the Aurora SDG Matching Initiative[1], many other initiatives were inspired and started as well, such as the *"Elsevier SDG Queries"* [5] in 2019, which focused on recall for the Times Higher Impact Ranking. Aurora has their own set of SDG queries with a focus on precision rather than recall. Also the *"Aurora SDG Queries"* has the capability to find literature on the narrower level of the targets, next to the broader level of the Goals. This allows us to identify researchers within the university network who work on research related to similar societal challenges of the narrower target level. With these SDG labeled research papers the data was enriched to provide information on the academic excellence and societal impact [6] using citation and altmetric databases, but that is out of scope for this report.

We created the initial version of the *"Aurora SDG Queries"* by simply using the significant keywords in the policy texts of the Sustainable Development Goals, Targets and Indicators[2], using boolean search operation structure in such a way (combining groups of concepts, synonyms and antonyms with near-operators), that the search results minimize

[1] https://aurora-network.global/project/sdg-analysis-bibliometrics-relevance/
[2] http://metadata.un.org/sdg/

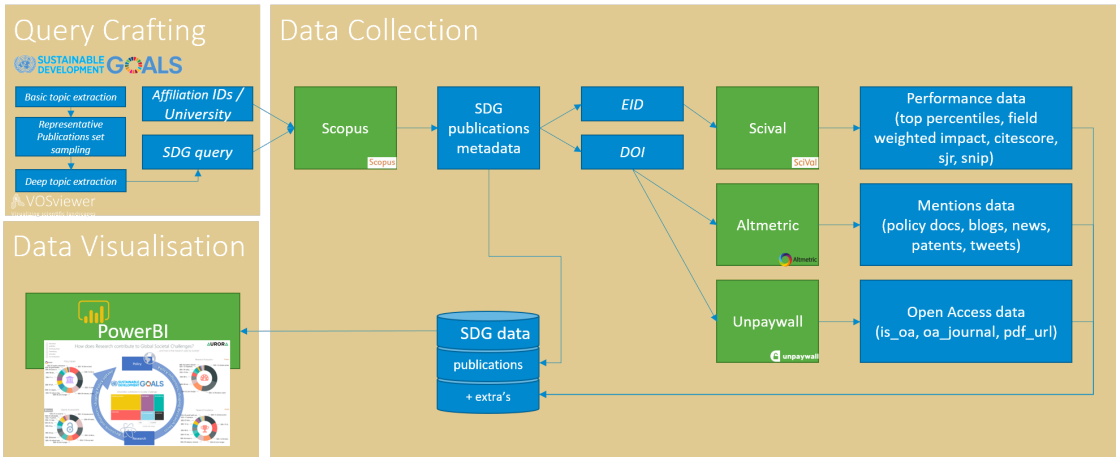


Figure 1: Workflow collecting data.

false positives. We incrementally improved the queries, and versioned them.[3] Version 1 was the initial version based on the exact words that appear in the policy texts. Following version 1 the results were reviewed by the group of bibliometricians each time a new version was released. Version 2 is the peer reviewed version. Version 3 we added concepts that were closely related to the exact keywords, like synonyms, antonyms and relevant keywords that occurred frequently in the search results. For version 4 [7] we split the search queries from the level of the Goal, to the narrower level of the Target, since it made much more sense for our use case. After version 4, we held a survey [8] among 250 Aurora Researchers to evaluate the precision and recall of the version 4 Aurora SDG queries, and to obtain suggestions for improvement for the next iteration and created a text analysis [9] on the survey data. In version 5 [10] the suggestions of the survey were processed.

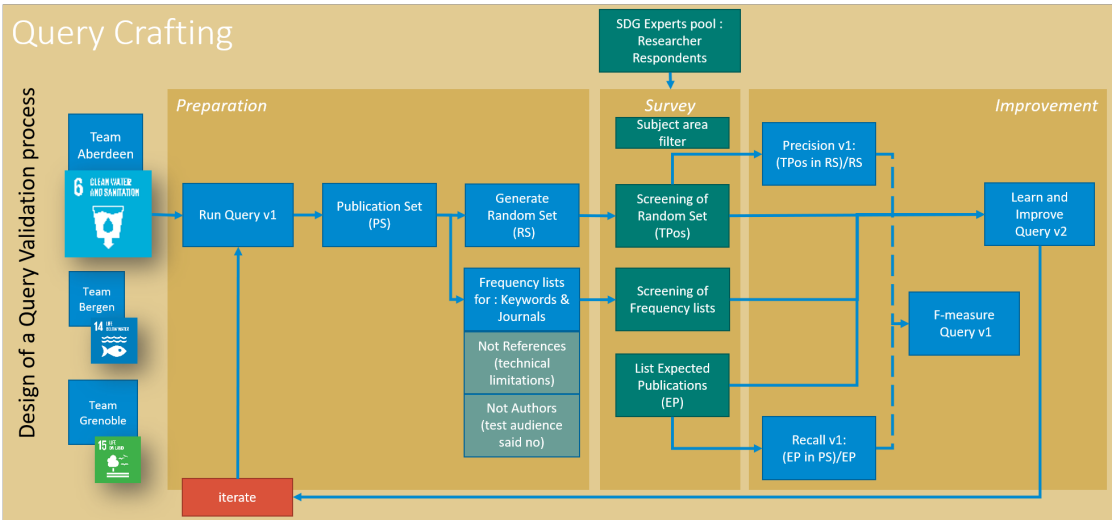


Figure 2: Workflow for improving Search Queries.

Following feedback from University leadership we were advised against using this type of survey and researcher resources in future iterations. This means we are unable to evaluate the precision and recall and make improvements to version 5 as we have done before and need to gain as much information as possible from the data already gathered, using mathematical linguistics models in the future. Therefore we need to make sure we know and evaluate the robustness of the current version 5 of the Aurora SDG Queries.

[3] https://github.com/Aurora-Network-Global/sdg-queries/releases

2 Research question

This report will focus on evaluation of the "Aurora SDG Queries"[4]. We will go into details about the two latest versions of the "Aurora SDG Queries"; comparing version 4 with version 5, and where possible comparing the *Aurora SDG Queries* to the *Elsevier SDG Queries*. And benchmark both using the handpicked "golden" data from the survey.

The main research question we want to answer is:
What is the quality of the results from the Aurora SDG Queries version 5?

The precision and recall is a proven indicator in the Information Retrieval community for assessing the quality of a search result. Precision shows how well the publications in the results for the SDG search query represent publications that are relevant to that SDG. And recall shows if the search for that SDG is retrieving all the relevant publications that could possibly exist.

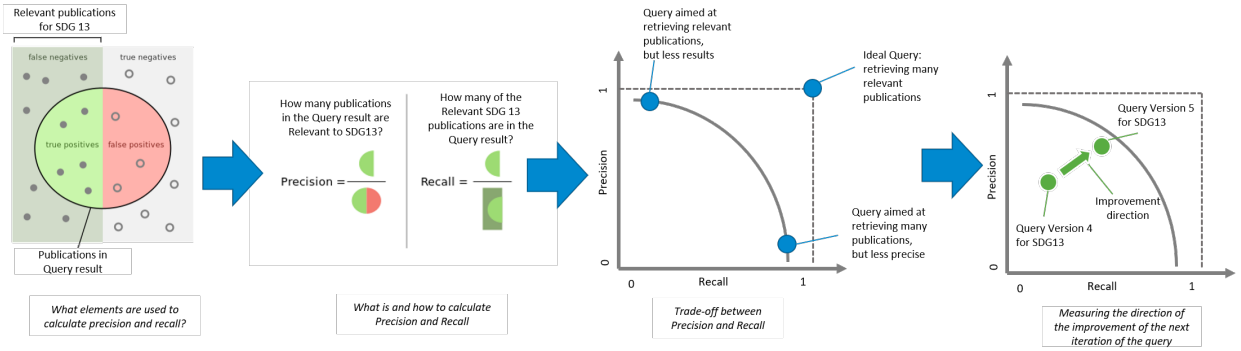


Figure 3: Explanation about precision and recall trade-offs and indicators for improvement.

To answer this question, we need *for recall* a baseline or "gold" data set with handpicked publications. The more of those publications appear in the research result, the better. This "gold" data we have from the suggested publication of the survey done earlier in 2020.[8]

Next we need *for precision* to count the publications from the SDG query result that are relevant. We have done this for the version 4 Aurora SDG queries using extensive resources from the Aurora research community; as this is no longer a viable option, we need to work our way around this problem.

We do this by first looking at how much the versions of the SDG queries differ from each other. Not only by size, but by comparing the publications inside each of the SDG collections.

Where the results from the versions do not deviate significantly, we believe the precision will not deviate significantly either.
However, where the results change a lot between the SDGs of the different versions, we need to re-establish the precision again. We do this by getting a random sample from the version 5 result of an SDG, and have that result evaluated by a bibliometrician familiar with the topic.

3 Sub-questions, Methods and Results

In this section we will explain the sub-questions, and why it is important to answer the main question. Along with that we explain the calculation method, and show the results.

Data collected for Aurora SDG query version 4 represent a different collection period, and different time ranges, from those collected from Scopus for version 5. (time ranges for v4: after 2009 before 2019, for v5: after 2009 and before 2020) This means, when comparing both versions, we have to keep in mind that there is a difference of publications in the result sets of roughly one year. We accepted this difference, because it is time consuming to collect all the data from Scopus.

[4] `https://aurora-network-global.github.io/sdg-queries/`

3.1 Volume of Publications in SDG result sets

To get a basic understanding of the data, we look at the volume and the number of publications in each of the versions of the Aurora SDG Queries and the Elsevier SDG Queries.

3.1.1 Amount of Publications per SDG

We want to know *What are the total number of publications per SDG in the Aurora SDG queries v4 and v5?* and *What are the total number of publications per SDG, in the Elsevier SDG queries 2020 and 2021?*

To get the total numbers we ran the SDG queries for each SDG target, and collected the Scopus electronic identifiers (EID's) that occur in each SDG target result sets. We deduplicated the EID's, when the targets for comparing each SDG Goal were combined. Then we counted the EID's in each of the SDG query result sets.

The total of the publications collected in the SDG result sets is 1,743,649 publications for the Aurora SDG queries version 4, and 1,555,477 for version 5.

In figure 4, we see the number of publications for each SDG result set ranges from 12,615 in SDG 17 of the Aurora version 4, to 473,190 in SDG 13 of version 5.

More detailed numbers for each SDG result set for each Query version can be found in table 2.

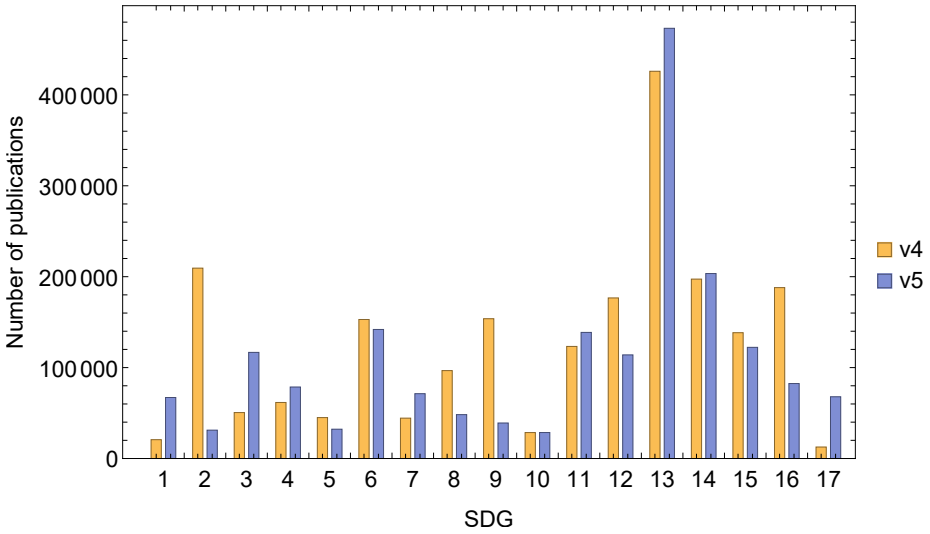


Figure 4: Number of results per SDG for v4 and v5 queries

We see there are some SDG result sets that differ a lot, when looking at the volume alone, but also within a result with comparable sizes, the publications occurring inside these sets might differ. More about the robustness can be read in section 3.4.

3.1.2 Amount of SDG labels per Publication

Next we want to know how well the publications in the SDG result sets overlap with other SDG result sets, or differ from each other. We need to know this to see how well the SDG result sets can be used for machine learning. The more the classes can be distinguished from each other, the better the model can be trained to distinguish between each of the classes later on in the process.

We want to know *How many and what percentage of the publications have only one SDG label, two SDG labels, etc. in the Aurora SDG queries v4 and v5?* and *How many and what percentage of the publications have only one SDG label, two SDG labels, etc. in the Elsevier SDG queries 2020 and 2021?*

Therefore we are looking at how many publications appear in one SDG result set, or appear in more than one.

In table 1 we can see that 82.5% of the publications in the Aurora SDG Queries version 4 occur in only one SDG result set, and 17.5% of the publications appear in two or more SDG result sets.
For the Aurora SDG Queries version 5 we can see that 84.2% of the publications occur in only one SDG result set, and

15.8% of the publications appear in two or more SDG result sets.
This means that the majority of the publications can be used for machine learning, which requires a large corpus for training, 1.3 million papers in case of the version 5 queries.

number of SDGs	number of publications (v4)	relative v4	number of publications (v5)	relative v5
1	1438591	82.5046	1310016	84.2196
2	244905	14.0455	199885	12.8504
3	46756	2.6815	37284	2.39695
4	10642	0.610329	6715	0.4317
5	2251	0.129097	1281	0.082
6	408	0.023	230	0.015
7	82	0.0047	50	0.0032
8	12	0.00069	12	0.00077
9	1	0.000057	2	0.00013
10	0	0	1	0.000064
11	1	0.000057	1	0.000064

Table 1: Number of publications classified to multiple SDGs for v4 and v5 queries (absolute and relative numbers).

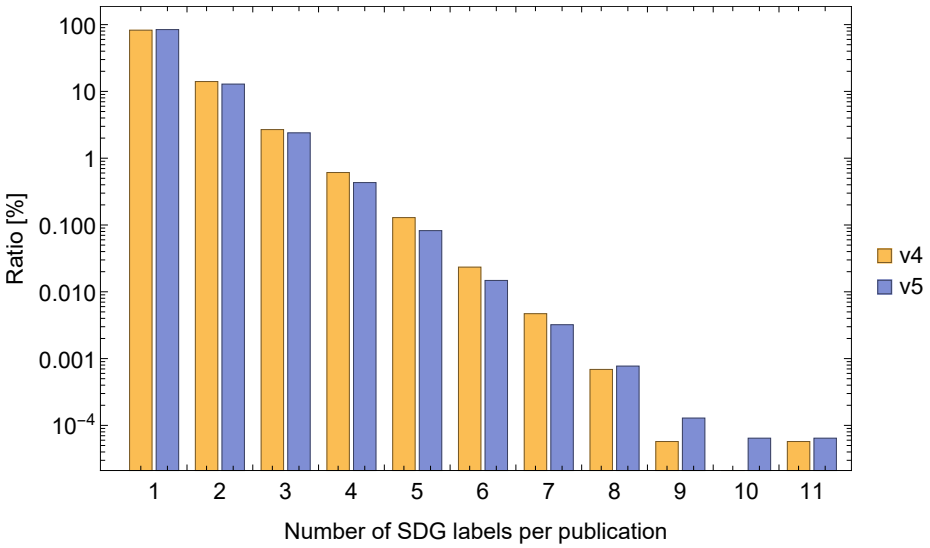


Figure 5: On logarithmic scale: Percentage of publications classified to one or multiple SDGs for v4 (orange) and v5 (purple) queries.

3.2 Overlap of publications between SDGs, within the same SDG query model

3.2.1 Overlap within Aurora SDG queries v4 and v5

In this section we look at the ∼20% of publications that appear in two or more SDG result sets. We now know that more than 80% of the publications occur in only one SDG result set. In this section we want to know how that ∼20% behave. We want to see what the effect is of the publications occurring in multiple SDGs.

We want to know *What is the overlap in numbers of publications of the result sets of the different queries of each SDG, within version 4 and within version 5 queries? Where do they relatively occur the most? What about the overlap between the targets?*

As we have seen before, the volumes of the publications in the SDG result sets differ in sizes. To compare the overlap between SDG result sets that have different sizes, we need to normalize both the SDG result sets we are comparing. For normalization we used this formula:

$$\text{normalized overlap}|_{i,j} = \frac{\text{number of publications in SDG } i \text{ AND j}}{\text{number of publications in SDG } i \text{ OR j}} \cdot 100\%$$

Here we divide the intersection over the union for the 289 (= 17x17) SDG result set combinations we are comparing. We calculate the intersection ("AND") for each combination of the SDG result sets i AND j that we are comparing. This gives us the number of publications that both sets share. We calculate the union ("OR") for each combination of the SDG result sets i OR j that we are comparing. This gives us the total number of publications in both sets together.

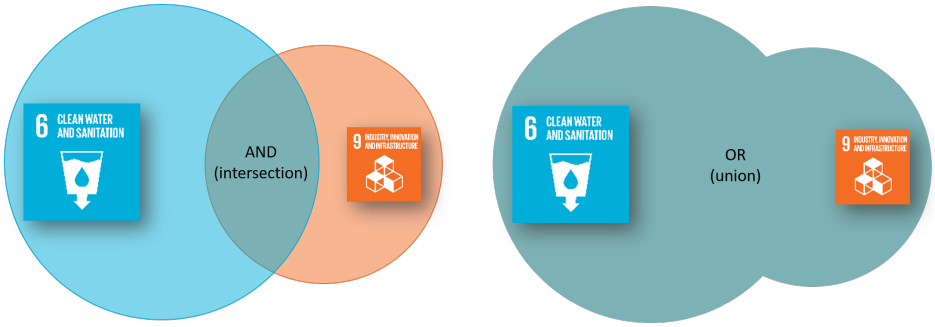
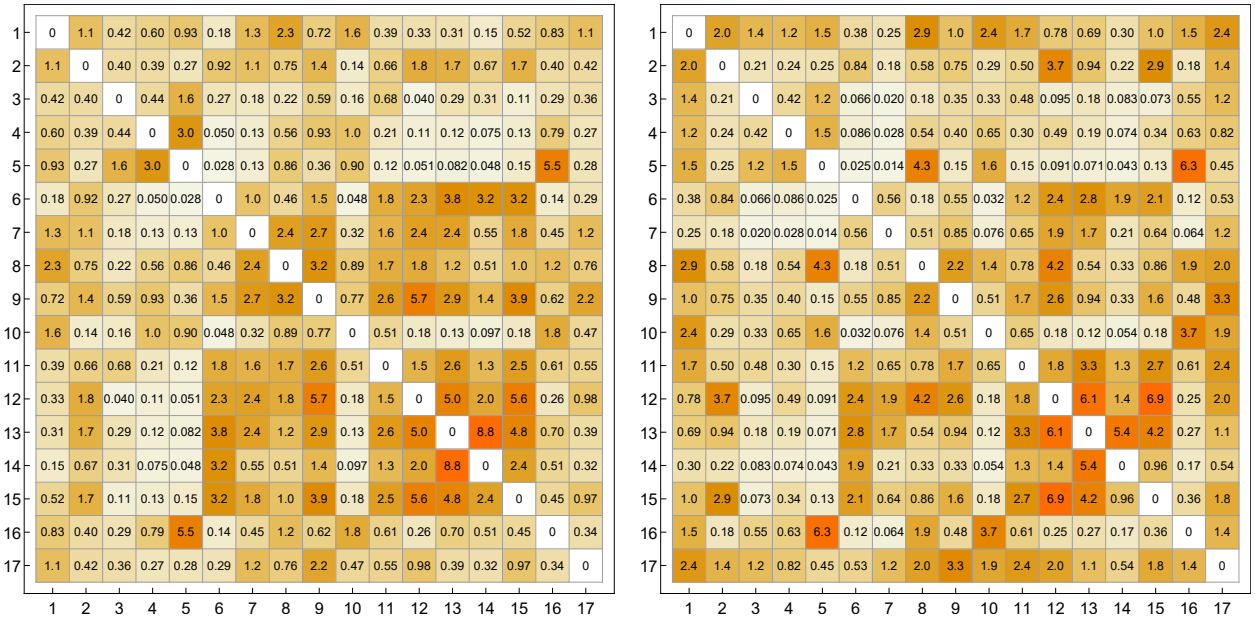


Figure 6: explanation of what is the union and intersection of the SDG result sets.

The result is a percentage, where 100% means that all publications in the SDG i result set will also appear in the SDG j result set. In the histograms below, you see values between 0.014% and 8.8%. The higher the number the more overlap between a crossing SDG.[5]

In figure 7a we see the heat-map for the overlap between different SDGs for the v4 queries. Here you can see that SDG 14 (life below water) and 13 (climate action) share relatively the most publications (8.8%) and SDG 6 (clean water and sanitation) and 5 (gender equality) the least overlap (0.028%).



(a) Overlap of Aurora SDG queries v4. (This represents 17.5% of publications labeled with more than one SDG.)

(b) Overlap of Aurora SDG queries v5. (This represents 15.8% of publications labeled with more than one SDG.)

Figure 7: Heat-map for the overlap (in %) between different SDGs.

[5]Note: We put a 0 on the diagonal line crossing SDG i with SDG i, where there actually should be 100. Because the number of publications that share different SDGs are so small, for proper color-grading the heat-map.

In figure 7b we see the heat-map for the overlap between different SDGs for the v5 queries. Here you can see that SDG 12 (responsible consumption and production) and 15 (life on land) share relatively the most publications (6.9%) and SDG 7 (clean energy) and 5 (gender equality) the least overlap (0.014%)

With a maximum overlap of 8.8% within the 15.8% of multi-labeled publications, we can conclude that the overlap between publication in the SDG result sets are very small. Inversely this means that the differences between the SDGs are big enough. This is helpful information that gives us confidence that we can define clear classes in the corpus of publications which is very useful for machine learning. This enables us to train a text-classifier that is able to distinguish a text on the level of these 17 SDG goals.

Overlap on the level of the targets within Aurora SDG queries v5 Next we want to look deeper in the 15.8% of the publications with overlap of the Aurora SDG queries v5, and see the how many of the targets show overlap. We have created 170 sub-queries, one for each of the targets in the 17 goals. We want to know how many of the publications in the result set for target 1.1 also appear in target 1.2, etc. We also look at the targets outside the parent SDG goal, where we want to know how many of the publications in the result set for target 1.1 appear in target 13.4, etc. This makes up to 14,365 combinations for the overlap between the targets.

We want to know how big the overlap is of the publications on the level of the targets. We calculated this by looking at the percentage of publications that show overlap between SDG target $i.x$ and SDG target $j.y$. Then we created buckets of 0.1%, to see the percentage of the 14,365 target combinations that share publications 0% to 0.1% overlap, from 0.1% to 0.2%, etc. We limited the buckets to 5%.

In figure 8 we see a histogram of the target sub-queries. Here we see that 79.8% of the 14365 possible combinations of the sub-queries have an overlap of less than 0.1%. Then the overlap drops drastically; 8.7% of the publications showing an overlap between 0.1 and 0.2%.

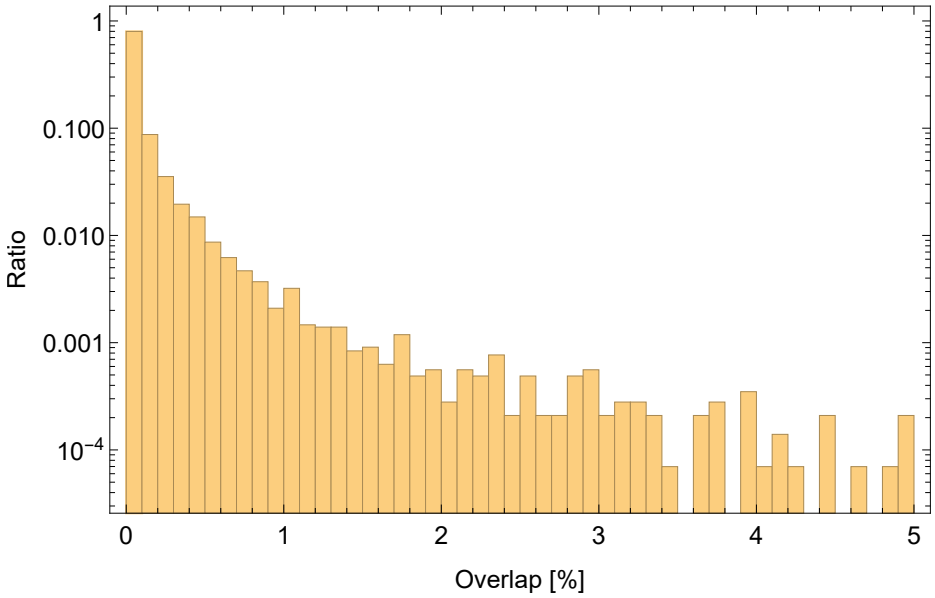


Figure 8: Histogram of the target sub-queries. 79.8% of the 14365 possible combinations of the sub-queries have an overlap of less than 0.1%.

Next, we want to know in more detail, which targets overlap the most and why.

In figure 9 we see the heat-map with 14,365 combinations for the overlap between the 170 Targets within the 17 SDGs for the v5 queries.

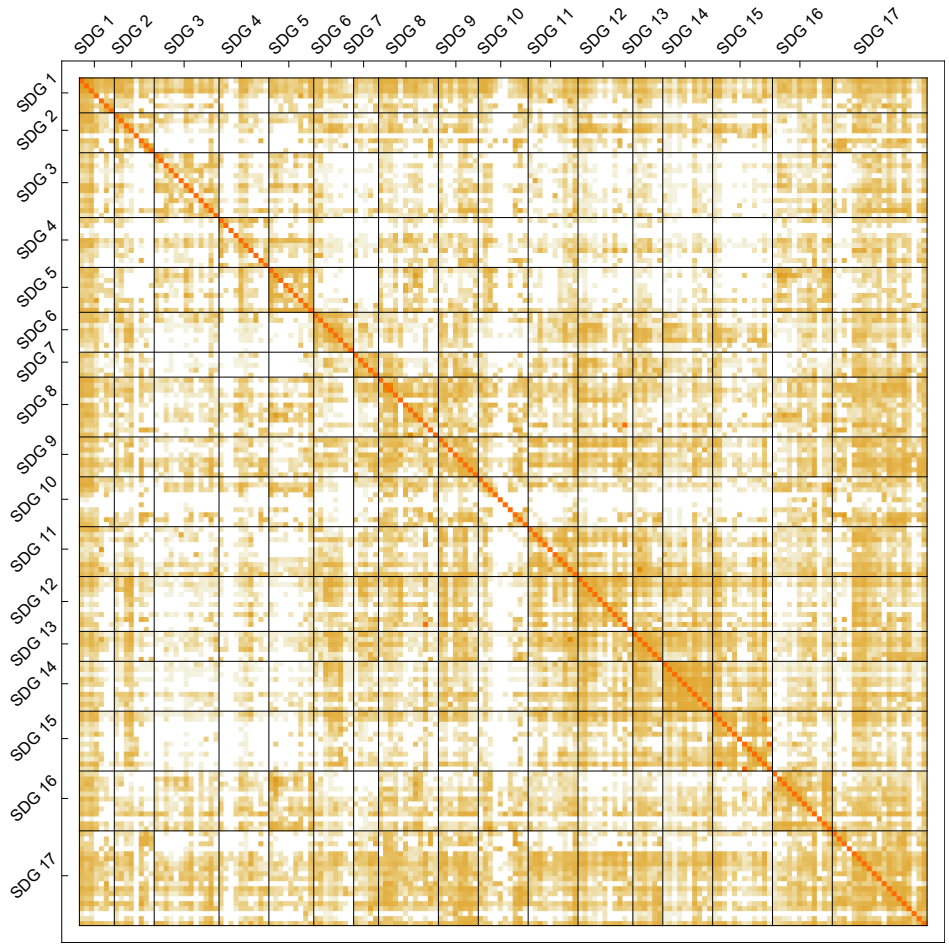


Figure 9: Heatmap for the overlap between different Targets of the SDGs for v5 queries.

Here in figure 9 you can see the dark orange dots on the diagonal lines, this shows that the greatest overlap of publications occurs within the same target. In the rest of the diagram we see a field of orange colors. We see that the targets within SDG 14 show relatively the most overlap. There are also some bright dark orange dots in the more remote areas of the diagonal line. Zooming in to one of those we see that for targets outside the same SDG targets 8.9 and 12.b share almost identical publications.

When we investigate the sub-queries for those targets, we can see that they share a query line that is identical, which probably takes account for the majority of the publications in that result set.

Sub-query for Target 8.9 in[6] Aurora SDG queries v5:

```
TITLE-ABS-KEY(("sustaina*") W/3 ("tourism")) OR
TITLE-ABS-KEY(("GDP" OR "Gross Domestic Product") W/3 ("tourism")) OR
TITLE-ABS-KEY(("job*") W/3 ("tourism"))
```

Sub-query for Target 12.b in[7] Aurora SDG queries v5:

```
TITLE-ABS-KEY((("sustainab*") W/3 ("tourism*")))
```

We can conclude that the overlap between publications in the target sub-query result sets are very small. This means that the differences between the SDGs are big enough, so we have clear classes in the corpus of publications that is very useful for machine learning. This enables us to train a text-classifier that is able to distinguish a text on the level of these 17 SDG Goals.

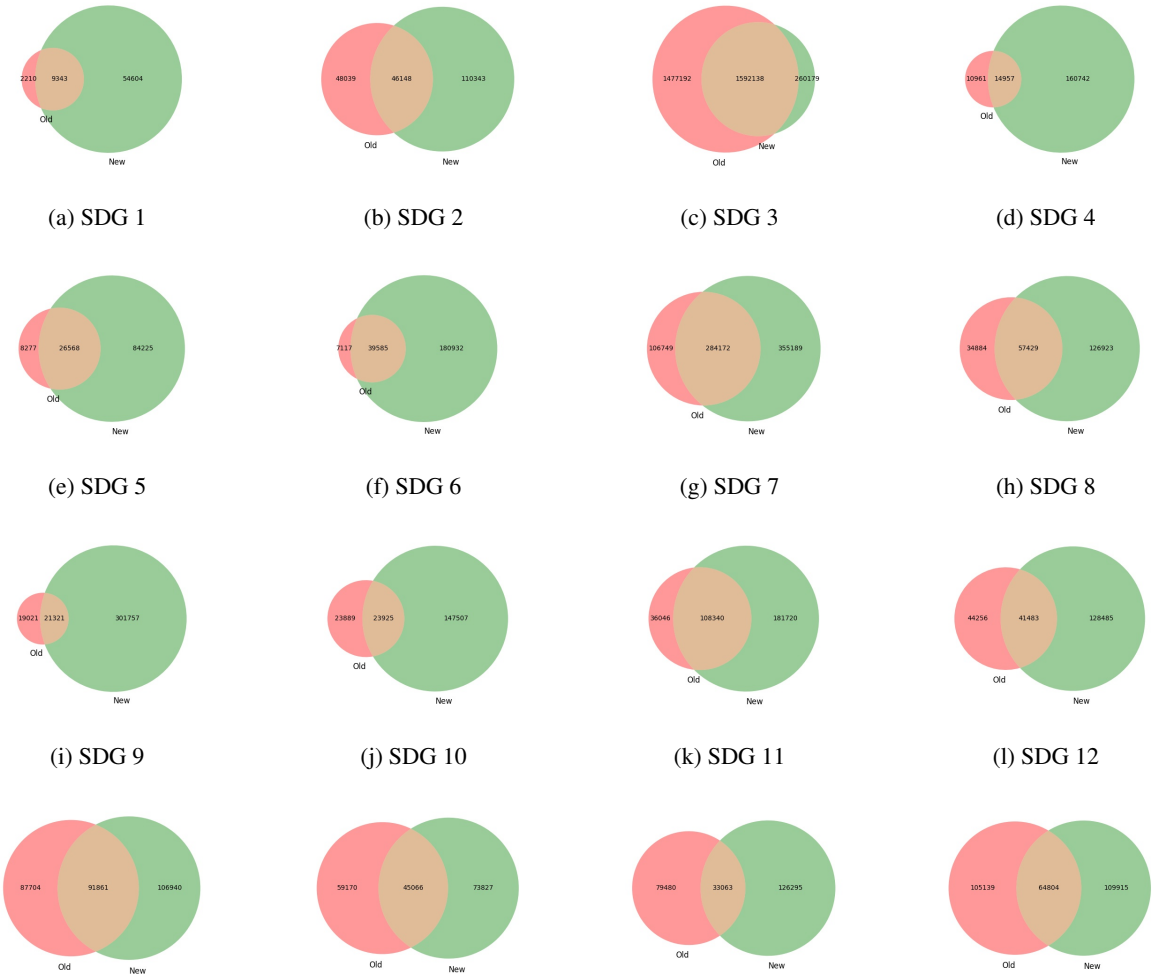
[6]Source: `https://aurora-network-global.github.io/sdg-queries/query_SDG8.xml`
[7]Source: `https://aurora-network-global.github.io/sdg-queries/query_SDG12.xml`

3.2.2 Elsevier SDG queries v2020 vs. v2021

The following section focuses a bit on the Elsevier queries, where we have a partnership with their teams. In 2019 Elsevier, lead by Bamini Jayabalasingham, created the initial version of the SDG queries version 2020 as an assignment by Times Higher Education as part of their new Impact Ranking. [5] The 2021 version is a different version, created by Science Metrix, lead by Maxime Rivest. [11] This version is different because the starting points were mini-queries used in another use case for UNESCO. They improved the queries using the Aurora SDG queries, and extended the recall by using a citation graph model.

We want to know the overlap in numbers of publications of the result sets of the different queries of each SDG, between Elsevier SDG queries[8] version 2020 and version 2021?

Science Metrix calculated the difference between the result sets and plotted them in venn diagrams.



(a) SDG 1 (b) SDG 2 (c) SDG 3 (d) SDG 4

(e) SDG 5 (f) SDG 6 (g) SDG 7 (h) SDG 8

(i) SDG 9 (j) SDG 10 (k) SDG 11 (l) SDG 12

(m) SDG 13 (n) SDG 14 (o) SDG 15 (p) SDG 16

Figure 10: Differences between Elsevier SDG queries version 2020 (red) and version 2021 (green)

3.3 Differences between query models: Aurora SDG queries v5 vs. Elsevier SDG queries v2021

Next we want to know what the difference is between the latest SDG query models; the Elsevier SDG queries version 2021, and the Aurora SDG queries version 5.

[8]Elsevier SDG Queries only account for 16 of the SDGs.

Science Metrix were generous in sharing their results here as well. Used by courtesy of Maxime Rivest (Science Metrix / Elsevier)

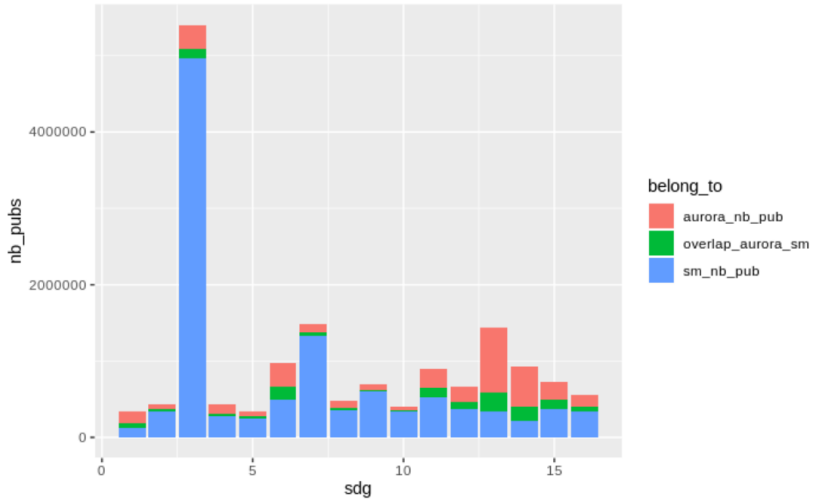


Figure 11: Bar diagrams for the overlap between different SDG query models. Aurora SDG queries v5 (in Red), Elsevier SDG queries 2021 (in Blue), Overlap (in Green)

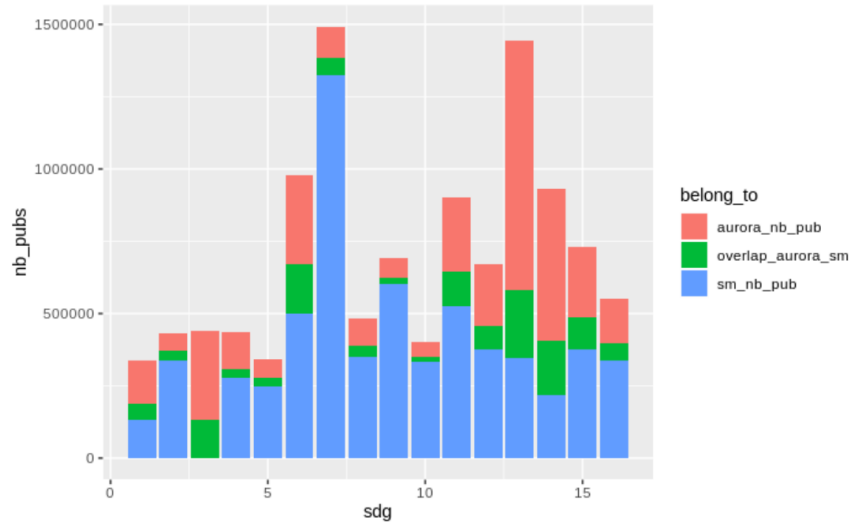


Figure 12: Bar diagrams for the overlap between different SDG query models, excluding Elsevier's SDG3. Aurora SDG queries v5 (in Red), Elsevier SDG queries 2021 (in Blue), Overlap (in Green)

We see in these diagrams that there is overlap between the models, but that overlap is very small, and the result sets from Elsevier are broader than those from Aurora. This is due to the fact that both models are created with a different use case in mind.

The Aurora SDG queries aim at precision of the search results, first of all in order to minimise false positives for the use cases to identify researchers in the various Aurora universities institutes related to an SDG and each of the targets. And secondly to create a corpus that can be used for machine learning, to expand the SDG labeling mechanism for research in the local languages of the Aurora universities without re-creating all the queries for the various languages.

The Elsevier SDG queries serve the use case for ranking universities to the Times Higher Education Impact Ranking. Here we can imagine more recall is favored over precision, in order to create a substantial corpus to have the majority of the universities worldwide represented in the ranking.

As being said by Rafols et al. 2021 [12] different stakeholders have different perspectives and needs regarding mapping research papers to the SDGs, and those should be given tools by interactively allowing to make a custom mapping[9].

3.4 Robustness: Change ratio within versions v4 and v5

To get a basic understanding about how the queries changed from v4 to v5 we take a look at the overlap between the results per SDG. Therefore, we calculate the number of publications that are assigned to a SDG for v4 and v5 queries (cardinalities), the number of publications that are assigned to a SDG by both versions of the queries (complements), the number of publications that are assigned only by one version of the queries (intersection) and the number of publications that are assigned to an SDG either by v4 or by v5 (union). The results are depicted in table 2.

SDG	v4	v5	v4 AND v5	v4 OR v5	v4 AND NOT v5	v5 AND NOT v4
1	20706	67112	69234	18584	2122	48528
2	209366	31142	223205	17303	192063	13839
3	50533	116711	129729	37515	13018	79196
4	61644	78595	83424	56815	4829	21780
5	45013	32205	58773	18445	26568	13760
6	152917	141939	234176	60680	92237	81259
7	44370	71294	112219	3445	40925	67849
8	96695	48266	122169	22792	73903	25474
9	153758	39013	178146	14625	139133	24388
10	28540	28573	46027	11086	17454	17487
11	123278	138698	172822	89154	34124	49544
12	176610	113947	240985	49572	127038	64375
13	425933	473190	508246	390877	35056	82313
14	197303	203463	254179	146587	50716	56876
15	138378	122252	172419	88211	50167	34041
16	187971	82465	240962	29474	158497	52991
17	12615	67903	74710	5808	6807	62095
all	1743649	1555477	2323166	975960	767689	579517

Table 2: Number of publications per SDG: v4, v5, intersection between v4 and v5, union of v4 and v5, only in v4, and only in v5.

Figures 13 and 14 show the numbers as stacked bar charts in absolute and relative numbers, respectively. The results only in v4 are shown in orange, the ones only in v5 in blue and the ones in both versions in brown. Additionally, we calculated the relative change of the cardinality from v4 to v5 and the robustness, which are defined as:

$$\text{change of cardinality} = \frac{\text{number of results from v5} - \text{number of results from v4}}{\text{number of results from v4}} \cdot 100\%$$

and

$$\text{robustness} = \frac{\text{number of publications in v4 and v5}}{\text{number of publications in v4 or v5}} \cdot 100\%$$

The change of cardinality gives a measure for how the number of results changes from v4 to v5, the robustness a measure for the shift of the results to different publications.

[9] https://public.tableau.com/profile/ed.noyons#!/vizhome/UKStringsSDGtocommunities/Dashboard1

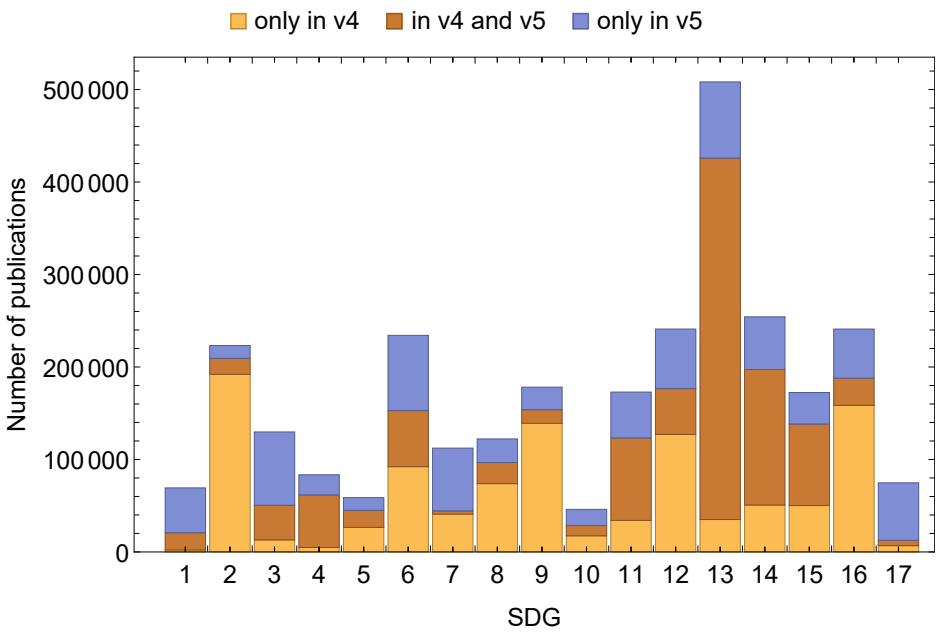


Figure 13: Overlap between v4 and v5 per SDG in absolute numbers.

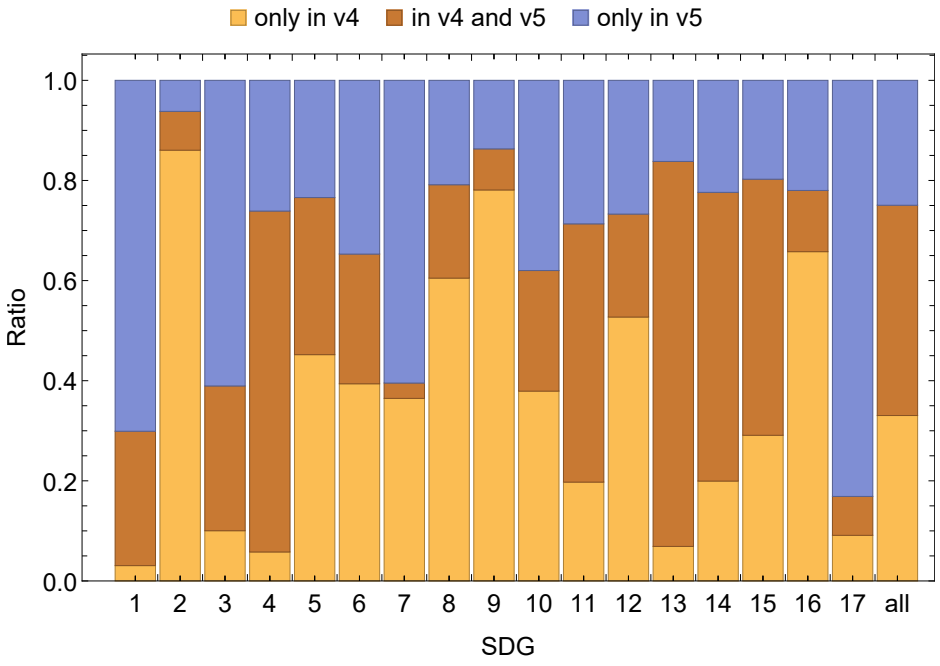


Figure 14: Overlap between v4 and v5 per SDG in relative numbers (normalized to the number of results in v4 OR v5 per SDG).

We identified four SDGs where the change from v4 to v5 was very large (robustness smaller than 10%): these are SDG02, SDG07, SDG09, and SDG17. In the next section we take a closer look at these SDGs and try to investigate the changes of the corresponding queries from a different point of view.

SDG	change of cardinality [%]	robustness [%]
1	224.1	26.8
2	-85.1	7.8
3	131.	28.9
4	27.5	68.1
5	-28.5	31.4
6	-7.2	25.9
7	60.7	3.1
8	-50.1	18.7
9	-74.6	8.2
10	0.1	24.1
11	12.5	51.6
12	-35.5	20.6
13	11.1	76.9
14	3.1	57.7
15	-11.7	51.2
16	-56.1	12.2
17	438.3	7.8
all	-10.8	42.

Table 3: Changes of results from v4 queries to v5 queries: relative change of the number of publications (cardinality) and shift of the results (robustness).

3.5 Precision and recall with baseline data

The aim in developing the new queries v5 from the v4 queries is to improve the results one gets. A way to measure this improvement can be done by looking at the precision and recall of the queries. The precision is the fraction of relevant papers among the results, the recall is the fraction of relevant papers that were found by the query.

3.5.1 Baseline data

For the calculations, we use the results from a comprehensive survey conducted among researchers based on the v4 queries [8] and a basic survey among project members concerning the four SDGs with low robustness (cf. table 3). In the comprehensive survey, 10793 publications were checked by researchers regarding the correct classification to a specific SDG. Additionally, 4106 publications were suggested, that should be in the respective result sets.

In the basic survey among project members, for each of the SDGs with a robustness below 10% we randomly selected 100 publications, and checked these manually regarding the correct classification.

3.5.2 Precision

We first calculate the precision from the basic survey for the four SDGs 02, 07, 09, and 17. The results are shown in table 4.

SDG	y	u	n	total	y [%]	u [%]	n [%]
2	85	2	13	100	85.0	2.0	13.0
7	85	13.	2	100	85.0	13.0	2.0
9	51	5	44	100	51.0	5.0	44.0
17	63	13	24	100	63.0	13.0	24.0

Table 4: Precision of the four SDGs with the lowest robustness (v5 queries).

SDG	y	u	n	total	y [%]	u [%]	n [%]
2	48	0	3	51	94.1	0.	5.9
7	7	0	0	7	100.0	0.0	0.0
9	16	1	15	32	50.0	3.1	46.9
17	12	1	2	15	80.0	6.7	13.3

Table 5: Precision of the four SDGs with the lowest robustness (v4 queries).

From the same data one can also calculate the precision of the v4 queries. The results are shown in table 5. Note, that these results are less reliable, due to the small sample size, especially for SDG07 and SDG17. From the numbers in tables 4 and 5 one can infer no improvement of the precision for the four SDGs with low robustness. The precision can also be calculated based on the old comprehensive survey among researchers. More publications were classified so the results should be more valid. Additionally, there are results for the other SDGs. The results are represented in tables 6 and 7 for the v4 and v5 queries, respectively. Again, the results are more reliable for the version of the queries, which

the survey was based on. In this case we have more data for v4 as the survey was conducted with publications from the v4 queries.

SDG	y	n	total	y [%]	n [%]
1	9	0	9	100.00	0.00
2	186	414	600	31.00	69.00
3	1878	745	2623	71.60	28.40
4	620	411	1031	60.14	39.86
5	456	140	596	76.51	23.49
6	265	76	341	77.71	22.29
7	356	385	741	48.04	51.96
8	98	105	203	48.28	51.72
9	362	265	627	57.74	42.26
10	253	137	390	64.87	35.13
11	373	238	611	61.05	38.95
12	342	275	617	55.43	44.57
13	367	201	568	64.61	35.39
14	164	61	225	72.89	27.11
15	484	204	688	70.35	29.65
16	386	146	532	72.56	27.44
17	141	250	391	36.06	63.94

Table 6: Precision (=y [%]) per SDG for v4 queries, based on the comprehensive survey.
v4 precision: Mean=62.87,
Standard deviation=16.54

SDG	y	n	total	y [%]	n [%]
1	9	0	9	100.00	0.00
2	33	19.	52	63.46	36.54
3	1405	510	1915	73.37	26.63
4	552	374	926	59.61	40.39
5	208	60	268	77.61	22.39
6	119	29	148	80.41	19.59
7	60	15	75	80.00	20.00
8	17	25	42	40.48	59.52
9	41	17	58	70.69	29.31
10	116	37	153	75.82	24.18
11	279	159	438	63.70	36.30
12	128	45	173	73.99	26.01
13	347	183	530	65.47	34.53
14	145	35	180	80.56	19.44
15	294	137	431	68.21	31.79
16	65	22	87	74.71	25.29
17	71	95	166	42.77	57.23

Table 7: Precision (=y [%]) per SDG for v5 queries, based on the comprehensive survey.
v5 precision: Mean=70.05,
Standard deviation=14.12

The differences of the precision scores can be inferred from the bar chart in figure 15. For 13 out of the 17 SDGs there is an (minor) improvement of the precision, 3 SDGs show a (minor) decrease, for one SDG it does not change. One should keep in mind that at this point the data for the v5 queries is rather limited. We find lowest precision scores from the results of v5 queries for SDG 9 (51%, based on the basic survey) and SDG 8 (40%, based on the comprehensive survey). We find major improvements of the precision for SDGs 2 and 7 (from 31% to 61% and from 48% to 80%, respectively); on average, the precision increased from 63% to 70%.

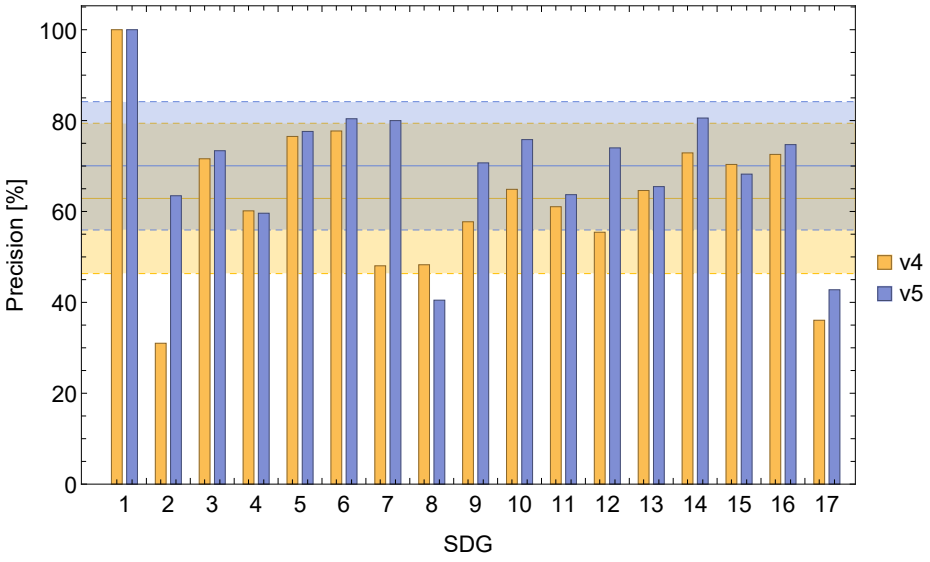


Figure 15: Precision v4 vs v5, based on the comprehensive survey among researchers.
v4: mean=62.87 standard deviation=16.54
v5: mean=70.05 standard deviation=14.12

3.5.3 Recall

The recall of the two versions of the query can be calculated from the publications suggested by the researchers in the comprehensive survey. The numbers can be found in table 8. Overall, 4106 publications were suggested, the distribution of these suggestions is very skewed and ranges from 5 for SDG 1 to 1255 for SDG 3.

SDG	suggested publications	v4	v5	v4 [%]	v5 [%]
1	5	0	1	0.00	20.00
2	89	30	22	33.71	24.72
3	1255	9	23	0.72	1.83
4	127	8	17	6.30	13.39
5	493	76	37	15.42	7.51
6	543	11	12	2.03	2.21
7	373	20	14	5.36	3.75
8	83	6	6	7.23	7.23
9	246	20	12	8.13	4.88
10	73	2	3	2.74	4.11
11	121	12	15	9.92	12.40
12	132	24	22	18.18	16.67
13	94	50	55	53.19	58.51
14	57	25	23	43.86	40.35
15	282	28	31	9.93	10.99
16	80	18	5	22.50	6.25
17	53	5	7	9.43	13.21
all	4106	344	305	8.38	7.43

Table 8: Recall per SDG, based on the comprehensive survey.
v4 recall: Mean=14.63, Standard deviation=15.44
v5 recall: Mean=14.59, Standard deviation=14.92

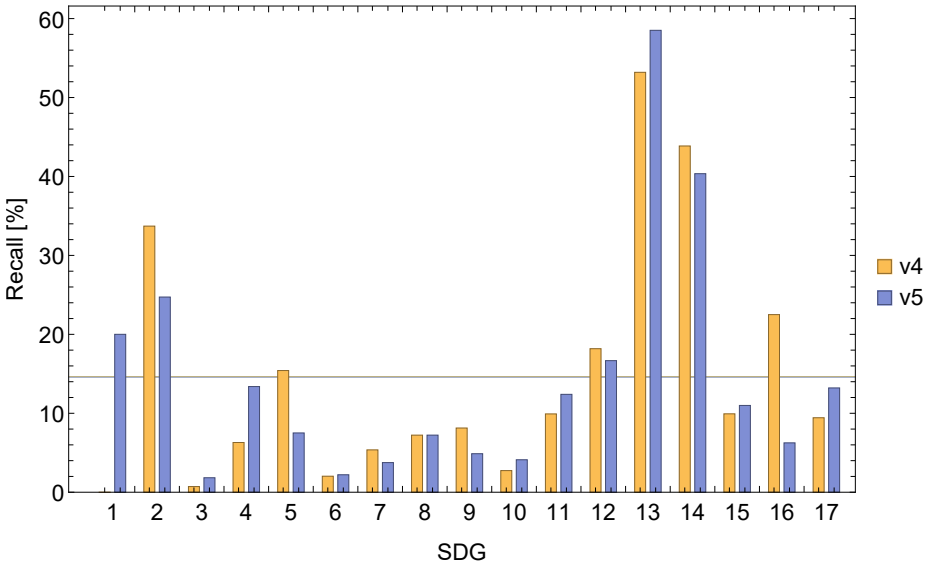
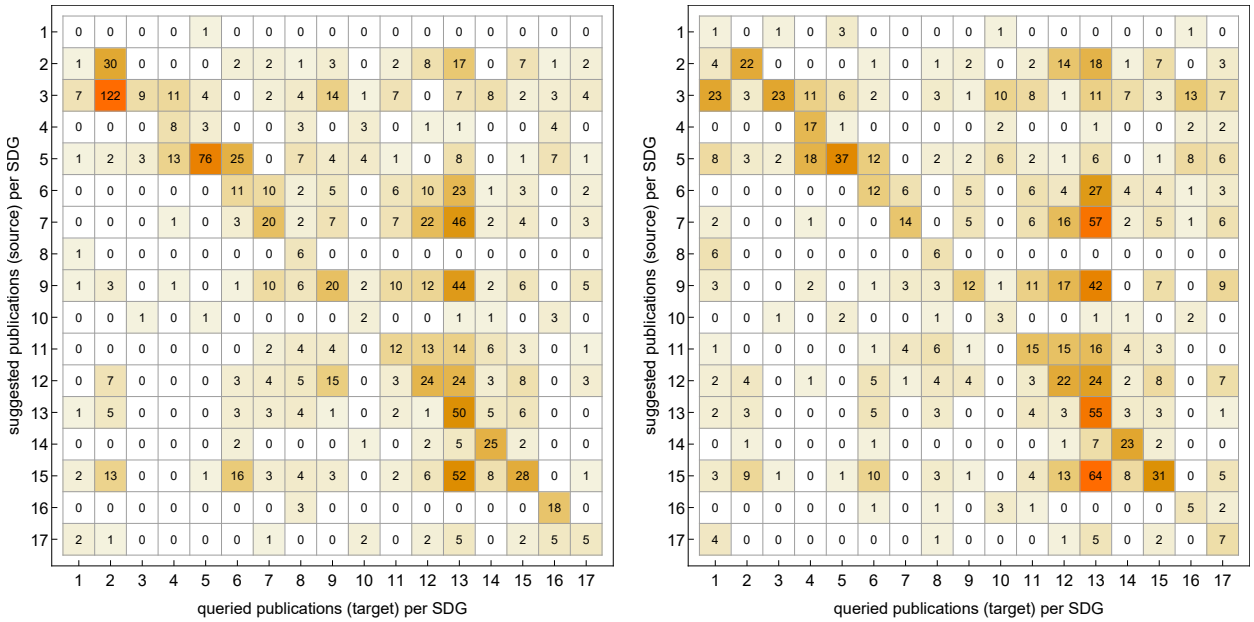


Figure 16: Recall v4 vs v5, based on the comprehensive survey among researchers. The solid lines represent the respective mean values.



(a) v4 (b) v5

Figure 17: Recall data for v4 and v5 SDG queries: Number of suggested publications in the surveyed SDG (source SDG, on vertical axis), that appear in the queried SDG results sets (target SDG, on horizontal axis). The entry in row i and column j gives the number of publications suggested by the researchers for SDG i, that were classified by our search queries to be in SDG j.

In figure 17b we can see a bright coloured column of SDG 13. This means that the papers in the query results of SDG 13 are appearing in lots of the suggested papers from the respondents, even more in the SDG 7 and SDG 5 suggested papers. This is an indication that the queries for SDG 13 are defined more broadly than needed, or the topic of SDG 13 (climate action) is much more interconnected with the research on the other SDGs like SDG 15 (life on land) and SDG 7 (clean energy). The other way around, looking at the rows we don't see a lot of bright coloured cells in that direction, which indicates the researchers have suggested papers that are more precisely relevant to the SDG they were expert in. Looking at the interesting rows, the suggested papers for SDG 3 (health and well being) are not only captured in SDG 3 itself, but the same figure appears also in SDG 1 (zero hunger). It might look a bit worrying that the query for SDG 1 captures most of the papers related to SDG 3, but we need to remember that the responses for SDG 1 are 5 suggested papers, compared to the 1255 suggested papers for SDG 3. In that perspective the version 5 queries for SDG 3 have much improved compared to version 4, in figure 17a where the queries for SDG 2 (no poverty) were capturing 122 papers that were actually suggested to be related to SDG 3.

3.5.4 Accuracy: Combining precision and recall

In this section we combine the outcomes of the precision and recall, which gives a measure of the accuracy of the different queries for each of the SDGs. We use the F-score to calculate the accuracy:

$$\text{F-score} = \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

This results in the following findings:

SDG	v4			v5		
	recall	precision	F-score	recall	precision	F-score
1	0.00	100.00	0.00	20.00	100.00	33.33
2	33.71	31.00	32.30	24.72	63.46	35.58
3	0.72	71.60	1.42	1.83	73.37	3.58
4	6.30	60.14	11.40	13.39	59.61	21.86
5	15.42	76.51	25.66	7.51	77.61	13.69
6	2.03	77.71	3.95	2.21	80.41	4.30
7	5.36	48.04	9.65	3.75	80.00	7.17
8	7.23	48.28	12.57	7.23	40.48	12.27
9	8.13	57.74	14.25	4.88	70.69	9.13
10	2.74	64.87	5.26	4.11	75.82	7.80
11	9.92	61.05	17.06	12.40	63.70	20.75
12	18.18	55.43	27.38	16.67	73.99	27.21
13	53.19	64.61	58.35	58.51	65.47	61.80
14	43.86	72.89	54.77	40.35	80.56	53.77
15	9.93	70.35	17.40	10.99	68.21	18.93
16	22.50	72.56	34.35	6.25	74.71	11.54
17	9.43	36.06	14.96	13.21	42.77	20.18
mean	14.63	62.87	20.04	14.59	70.05	21.34
standard deviation	15.44	16.54	17.05	14.92	14.12	16.68

Table 9: Precision, recall and F-score for v4 and v5 queries.

Figure 18 shows the F-scores of the SDG queries of the different versions side-by-side.

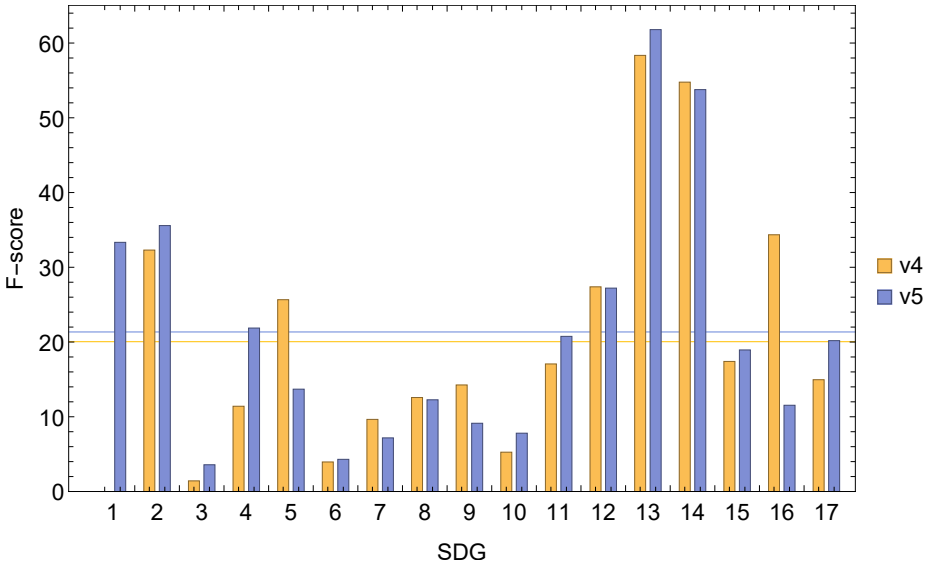


Figure 18: F-score v4 vs v5. The solid lines represent the respective mean values.

Calculating the F-scores for both versions, we learn that there is a small increase in accuracy when looking at the averages of all SDG's for each query version. F=20 for version 4, and F=21 for version 5. While the recall for both versions remain the same around 14, we see that on average there is an increase of the precision of 63 in version 4 to 70 in version 5. Looking at the individual SDG's we see that for SDGs 1, 2, 13, 14 and 16 we have higher F-scores compared to the other SDGs. We also see that some SDGs F-scores decreased from version 4 to version 5.

We want to know how or if that increase or decrease is related to the amount of change of the publications in the result sets of each SDG. We combined the change of F-score between v4 and v5 from table 9 with the robustness score of table 3, resulting in the following chart.

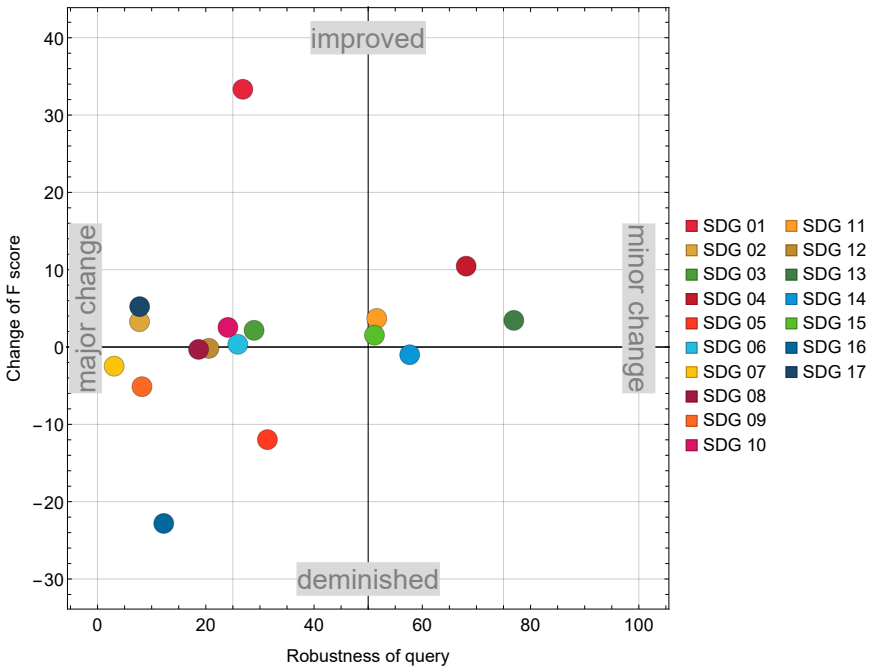


Figure 19: F-score change vs robustness of SDG queries: Dots more to the left shows that there are lots of new publications in an SDG result set in the version 5 queries, compared to the version 4. Dots more to the right are kept more or less the same. Dot more to top show that the accuracy of the SDG result set has improved in the version 5 queries, compared to version 4. Dots more to the bottom show that the accuracy got worse.

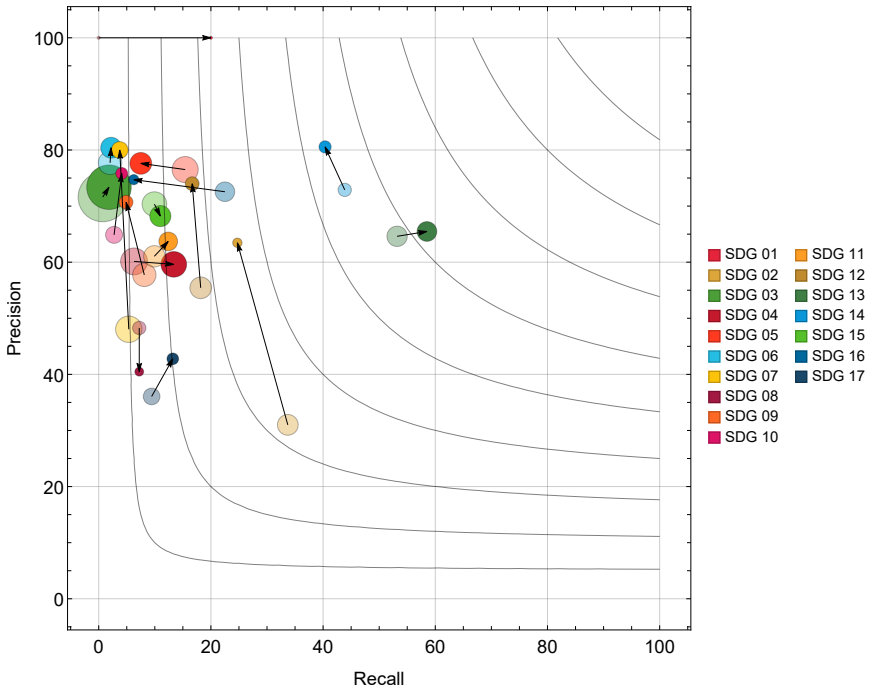


Figure 20: Precision, recall and amount of publications suggested and selected by expert researchers for v4 and v5 queries. V4 SDGs in semi-transparent color, v5 SDGs in full colors. The iso-lines show the constant F-scores from 5 (bottom left) to 45 (top right).

We see in figure 19 how the SDGs have improved or diminished in F-score in SDG queries version 5, and whether that required a small or big change of the publications in the SDG result sets to account for that effect.

F-scores don't distinguish between precision and recall. To show if the accuracy is based on precision or the recall, we made a bubble plot for each version v4 and v5, where we plot the precision, recall and the tested sample sizes for each of the SDGs.

Here we see that our queries mainly moved to increase the precision, without big changes in the F-scores, at the cost of a low recall.

This means *the Aurora SDG queries version 5 collect a small number of papers, but the papers collected are in the majority of the cases relevant publications related to the SDG goals and targets.*

4 Conclusions & Discussion

The accuracy of the query version v4 compared to version 5 increased a little, F=20 for version 4, and F=21 for version 5. While the recall for both versions remain the same around 14%, we see that on average there is an increase of the precision of 63% in version 4 to 70% in version 5.

Overall we can conclude that changing the queries from version 4 to version 5 isn't really a ground breaking change where we would have liked to see a major migration of the SDG's to the top right corner in figure 20. What we can conclude is that we have made some important corrections to the queries so that they are precise enough to match the research publications what would generally speaking be expected in the result sets. This is demonstrating that the queries are build for precision, rather than recall. The advantage with a high precision query model is that the research publications you'll get are most of the time related to that SDG. The disadvantage of a low recall query model is that the queries are missing out lots of relevant research publications that should be in the result set.

Having a labeled corpus aiming at precision of the labeling of the SDG goals, allows us to use the queries to generate a labeled corpus that can be used in machine learning [13]. To gain recall, we can feed the machine learning phase with a language model to label papers to an SDG that is closely related to similar semantic concepts. This is something we are unable to accomplish with boolean queries, due to the binary nature of a keyword search. This result gives confidence we have a solid foundation to build a machine learning model from the labeled corpus with the Aurora SDG queries version 5.

With IDfuse[10] we did some early experimentation on enhancing the results of the query based model with machine learning using the widely used BERT model, retrained for Scientific content (SciBERT)[11]. To train the SciBert model to label publications to the SDGs, IDfuse loaded the OpenAIRE Research Graph in to ElasticSearch, translated[12] the Aurora queries version 5 from a Scopus query language to the ElasticSearch query language, and used the result sets to train the model.

[10] IDfuse is a Dutch startup that uses machine learning to improve grant proposal writing. `https://idfuse.nl/`

[11] `https://github.com/allenai/scibert`

[12] Scopus to Elastic Search Translator: `https://github.com/martijnvanbeers/booque` for creating training corpus with OpenAIRE Research Graph

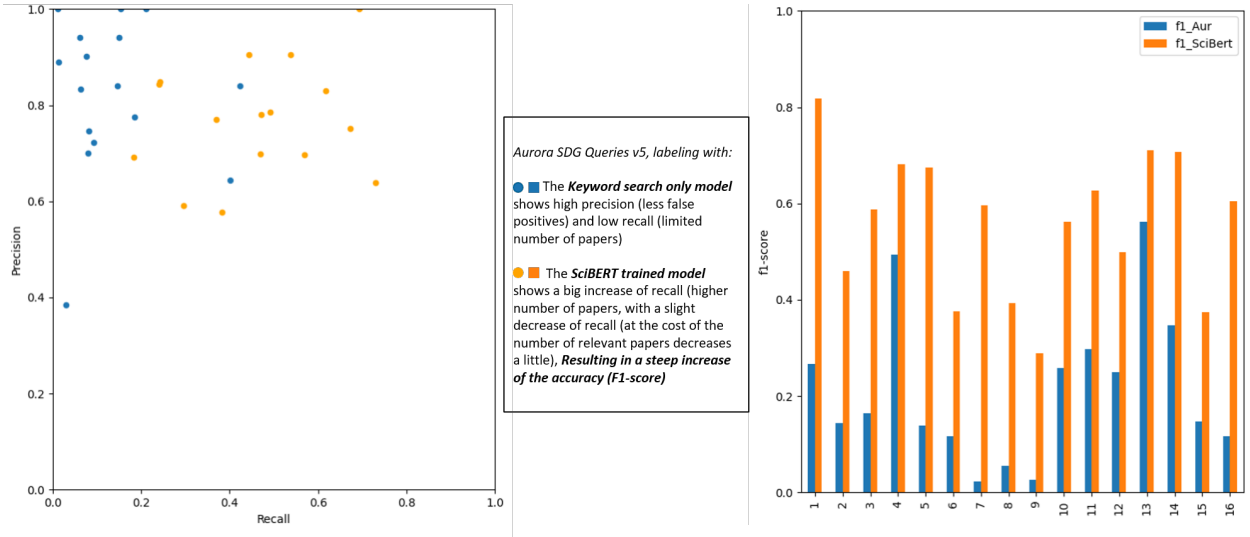


Figure 21: Training label sets with SciBERT language model improves the recall of the Boolean queries, at a cost of slight decrease of precision.

In figure 21 we see that the keyword-search-only model (blue dots and bars) shows high precision (less false positives) and low recall (limited number of papers). The SciBERT-trained model (yellow/orange dots and bars) shows on the left chart a big increase of recall, at a slight cost of precision. Results on the right chart show a steep increase of the accuracy (F-score).

This early experiment gives us confidence we can use such advanced language models to classify the SDGs, while increasing the number of research papers to be included in the result set for the SDGs, and that this larger result set will contain papers that are still relevant to the labeled SDGs.

Increasing recall, without sacrificing much of the precision was our first goal to include research in the SDGs that we could not capture using Boolean queries. The next step is if we can also capture research written in other languages than English to the SDGs. Fortunately BERT language models are capable of cross-language classifications while being trained in only one language. This is something we need to investigate further.

5 Acknowledgements

Thanks to your colleagues at University of Aberdeen for the excellent proof reading, and preventing us from making stupid mistakes.
Thanks to Maxime Rivest, Science Metrix, for sharing with us the graphs that he generated to compare the Aurora and Elsevier models, in our knowledge exchange partnership with Elsevier [13].

References

[1] Maurice Vanderfeesten and René Otten. Societal relevant impact : Potential analysis for aurora-network university leaders to strengthen collaboration on societal challenges. Aurora-Network Norwich 2017 (Aurora2017).

[2] L. van Drooge, P. van den Besselaar, G. M. F. Elsen, M. de Haas, J. J. van den Heuvel, H. Maassen van den Brink, B. van der Meulen, J. B. Spaapen, and R. Westenbrink. Evaluating the societal relevance of academic research: A guide. *EriC-Evaluating Research in Context*, 2010. Publisher: ERiC-Evaluating Research in Context.

[3] Michael J. Carley and Eduardo Bustelo. *Social Impact Assessment And Monitoring: A Guide To The Literature*. Routledge, 2019. Google-Books-ID: lqiaDwAAQBAJ.

[4] Caroline S. Armitage, Marta Lorenz, and Susanne Mikki. Mapping scholarly publications related to the sustainable development goals: Do independent bibliometric approaches get the same results? *Quantitative Science Studies*, 1(3):1092–1108, 2020.

[13] `https://www.elsevier.com/about/partnerships/sdg-research-mapping-initiative`

[5] Bamini Jayabalasingham, Roy Boverhof, Kevin Agnew, and L. Klein. Identifying research supporting the united nations sustainable development goals. *Mendeley Data*, 1, 2019. Publisher: Mendeley.

[6] Lutz Bornmann. What is societal impact of research and how can it be assessed? a literature survey. *Journal of the American Society for Information Science and Technology*, 64(2):217–233, 2013.

[7] Maurice Vanderfeesten, René Otten, and Eike Spielberg. Search queries for "mapping research output to the sustainable development goals (SDGs)" v4.0.

[8] Maurice Vanderfeesten, Eike Spielberg, and Yassin Gunes. Survey data of "mapping research output to the sustainable development goals (SDGs)". type: dataset.

[9] Maurice Vanderfeesten, Eike Spielberg, and Linda Hasse. Text analyses of survey data on "mapping research output to the sustainable development goals (SDGs)". type: dataset.

[10] Maurice Vanderfeesten, René Otten, and Eike Spielberg. Search queries for "mapping research output to the sustainable development goals (SDGs)" v5.0.2.

[11] Maxime Rivest, Yury Kashnitsky, Alexandre Bédard-Vallée, David Campbell, Paul Khayat, Isabelle Labrosse, Henrique Pinheiro, Simon Provençal, Guillaume Roberge, and Chris James. Improving the scopus and aurora queries to identify research that supports the united nations sustainable development goals (SDGs) 2021. *Mendeley*, 2, 2021. Publisher: Mendeley.

[12] Ismael Rafols, Ed Noyons, Hugo Confraria, and Tommaso Ciarli. Visualising plural mappings of science for sustainable development goals (SDGs). *SocArXiv*, 2021.

[13] Rui Zhang, Maéva Vignes, Ulrich Steiner, and Arthur Zimek. Matching research publications to the united nations' sustainable development goals by multi-label-learning with hierarchical categories. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 516–525, 2020.