Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

# Mining the History Sections of Wikipedia Articles

## Bachelor's Thesis

Wolfgang Kircheis
Born 14 April 1985 in Leipzig

Matriculation Number 3718447

1. Referee: Junior-Prof. Dr. Martin Potthast

Submission date: 14 February 2023

# Declaration/Erklärung

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work. Any citations, both verbatim and paraphrased, have been marked as such. I am aware that failure to comply with these stipulations may result in (retrospective) revocation of my degree. I declare that the electronic copy matches the printed copies.

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, 14 February 2023

.....
Wolfgang Kircheis

**Abstract**

This thesis presents a new corpus of science and technology Wikipedia articles and assesses methods for the retrieval of history sections within them. Both the corpus itself and the retrieval technology described in this thesis will assist science studies research science priority disputes by availing of Wikipedia's unique position as a community-driven, up-to-date, and traceable account of the debates surrounding the attribution of major scientific breakthroughs. Science and technology Wikipedia articles are mined from Wikitext dumps using iterative filtering of Wikipedia's category network. The thesis details the structure, extraction, segmentation, and breakdown of articles into sections, subsections and subsubsections. It explains and evaluates how a combination of heuristics analyzing section headings and classifiers trained on a ground truth of articles with designated history sections can be utilized to identify sections featuring the long-term historical development of technological innovations.

# Contents

# Chapter 1

# Introduction

When the 2020 Nobel Prize in Chemistry was awarded to Jennifer Doudna and Emmanuelle Charpentier for their contributions to the development of a method for genome editing [NobelPrize.org, 2022], biotechnology had already seen a long history of legal battles over patent rights as well as "accusations of impropriety" and "allegations of bad actors and bad faith": The team at UC Berkeley surrounding Doudna and Charpentier filed for a patent first, but the Broad Institute, led by biologist Feng Zhang, "opted for an expedited review process, and its patents were granted earlier", resulting in Berkeley claiming patent interference and "launching a complicated process to establish who first came up with the invention" [Ledford, 2016b]. Matters were further complicated in 2020 when the *Patent Trial and Appeal Board* ruled that the Broad Institute had priority over inventions not covered by Berkeley's patent [Cohen, 2020]. In the weeks leading up the 2020 Nobel Prize, some suspected Zhang would be added to the list of recipients, and he was not the only addition people expected to see on the list [Ball, 2021].

The science priority dispute surrounding CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)[1] illustrates how scientific discoveries and innovative technologies can give birth to disputes, and they are highly relevant to science studies. Researchers are interested in how these battles are fought and, more importantly, how claims of actors (scientists, institutes, or companies) can be linked to their research contributions, thus providing basis to verify them. Science studies rely on accounts to shed light on these claims. Accounts are reports of varying authority about actors, timelines, and priorities. On the one hand, contributions can be measured by analyzing the impact of the research conducted in a field over time (research accounts); on the other hand, science studies can look at more removed sources (meta accounts) by reviewing editorial and collaborative accounts (see Figure 1.1).

[1] "a family of DNA sequences found in the genomes of prokaryotic organisms such as bacteria and archaea" [Wikipedia contributors, 2022a]

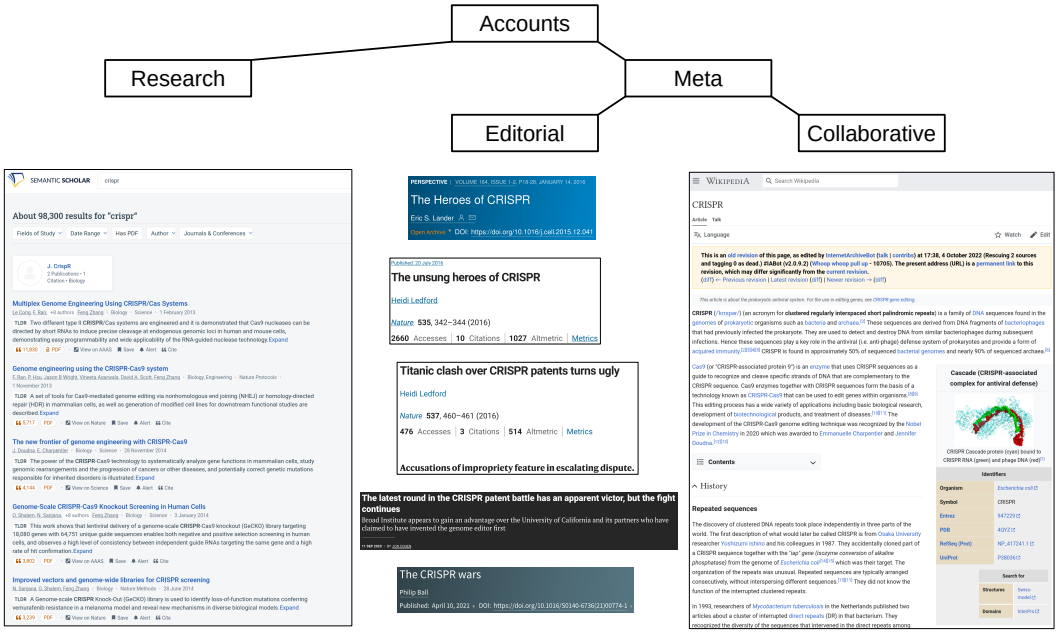



**Figure 1.1:** Taxonomy of science study accounts. [Ball, 2021, Cohen, 2020, Lander, 2016, Ledford, 2016a,b, Semantic Scholar, 2022, Wikipedia contributors, 2022a]

Analyzing research accounts by harvesting primary (e.g. research papers) and secondary (e.g. review papers) literature can be seen as the first step towards delineating the bibliometric field of research and trying to answer the question which publications are the ones most relevant to a particular field of study. A researcher interested in the scientific priority disputes surrounding CRISPR could query the Web of Science or Google Scholar for the keyword 'crispr' or related terms and sort, filter, and analyze the results. In addition to classical approaches of ready-made classifications like the Journal Impact Factor [Larivière and Sugimoto, 2019], a number of computational methods have been proposed to delineate a scientific field, such as keyword queries based on a seed of highly cited articles, various clustering methods combined with bibliometric mapping (coupling, chained citation, co-citation), citation and co-author networks, and hybrid models of all these approaches [Zitt et al., 2019].

Meta accounts can be divided into editorial and collaborative accounts. Editorial accounts can be news articles, reports in popular science magazines, editorial pieces in journals, as well as designated works detailing the history and development of a scientific field or technology, like Lander's 'The Heroes of CRISPR' [Lander, 2016] or Ledford's competing 'The Unsung Heroes of

CRISPR' [Ledford, 2016b]. Unlike editorial accounts with their unilateral communication, collaborative accounts are based on, created via, and maintained by multilateral communication. Wikipedia talk pages offer an insight into the editing process and contributors' decisions. More importantly, Wikipedia keeps track of almost all revisions of all articles available. Researchers therefore have access to a detailed timeline of historical snapshots of the current scientific debate surrounding a research field, backed by Wikipedia's policy of requiring contributors to provide secondary and tertiary sources for their claims [Wikipedia contributors, 2022c]. Most research in science studies has so far focused on research accounts (i.e. primary and secondary literature) and, to a lesser extent, editorial accounts (i.e. reports, journals, and designated publications), while collaborative accounts in general and Wikipedia in particular have been largely overlooked.

This thesis will explore how article sections covering the historical development of scientific and technological innovations ('history sections') can be mined from Wikipedia at scale to provide science studies with a tool to track the way the framing of priority claims has changed over time. A new corpus of science and technology articles is the basis for assessing the efficacy of both heuristic and computational classification approaches, a combination of which will prove to be the best strategy to identify history sections, which can then in turn be extended to the entire revision history of selected (science and technology) Wikipedia articles.

# Chapter 2

# Related Work

Wikipedia has been of interest as a resource both for science studies in general and for text extraction tasks in particular.

### Science Studies

Wikipedia has become a highly relevant tool used by researchers and educators and its quality is upheld by editors who ensure that articles adhere to the standards of the community [Nix, 2010].

Several authors have highlighted Wikipedia's unique selling point as "the world's largest collaboratively edited source of encyclopaedic knowledge" [Völkel et al., 2006], "the largest collection of freely available knowledge" [Zesch et al., 2008], and "a massive and relatively high-quality collection of text and (predominantly unstructured) encyclopedic knowledge" [Flickinger et al., 2010]. Wikipedia is a resource for understanding the historical development surrounding societal controversies as Wikipedia's inherent revision history and talk pages enable tracking and tracing any and all changes made to articles [Borra et al., 2015].

According to Lin et al. [2017] Wikipedia has become one of the most important resources for data extraction research and its practical applications. They also highlight its importance for Wikipedia-based studies and Wikipedia-based systems while addressing the *article-as-concept assumption*: An integral idea behind many Wikipedia studies and systems is that there is a direct mapping between concepts and the Wikipedia articles that describe them, which, being a misconception, leads to incorrect assumptions about articles describing the same concept in different languages, which in turn leads to "mistaken conclusions about [...] the similarities and differences in the articles about the same concept in different language editions".

Owing to its ever-growing size and the way it covers both past and present developments, Wikipedia is a new kind of lexical semantic resource, wherefore it

has been used in a variety of NLP tasks, such as text categorization, information extraction, information retrieval, question answering, computing semantic relatedness, and named entity recognition [Zesch et al., 2008]. Zesch compares Wikipedia to linguistic knowledge bases (LKBs) like WordNet and GermaNet, referring to Wikipedia and Wikitionary as instances of collaborative knowledge bases (CKBs), which provide information on named entities, phrases or terms specific to domains uncommon in LKBs, but also noting that – like many other CKBs – it lacks accessible APIs.

### Text Extraction

Wikipedia has been employed as a source for a variety of text extraction tasks, many of which are focused on article sections, as they "are the building blocks of Wikipedia articles" [Piccardi et al., 2018]. As about one quarter of all English language articles have only one or even no sections and the vast majority of headings are only ever used once, Piccardi et al. recommend (sub)sections for articles by finding sections from similar articles using topic modelling, collaborative filtering, and Wikipedia's category system, with category-based approaches being most successful.

WIKITABLET ('Wikipedia Tables to Text') matches up tabular and metadata in Wikipedia articles with their respective sections using a transformer as a base model [Chen et al., 2021].

Schenkel et al. [2007] extend Wikipedia dumps with "semantically rich, self-explaining tags" by exploiting Wikipedia's category network, which they claim is of high quality due to categories being assigned manually. However, this view is not shared by all authors, and Wikipedia's category network has been described as "noisy and ill-conceived" [Piccardi et al., 2018].

As many Wikipedia entries lack section subdivision and have inconsistent headings, Field et al. [2020] generate section titles for Wikipedia articles with BERT-based uncased encoders and RNN decoders.

Liu and Iwaihara [2016] extract representative phrases for sections from external articles containing the same words as the target article. They retrieve candidate articles by calculating the TF-IDF-based cosine-similarity between related articles and each section (using Latent Dirichlet Allocation to assess word-topic distributions and boost sections) and use FP-growth to extract co-occurring word sets, then pipe phrases into search engines and apply gradient descent to rank them.

Aprosio and Tonelli [2015] record a growing interest in the task of extracting biographical information from data and name Wikipedia "the main source of information for research in this direction despite its many biases". Seeing as Wikipedia's lack of consistent templates for describing biographies has led to

various page types to describe a person's life, they employ Conditional Random Fields (CRFsuite) and compare them to Support Vector Machines (YAMCHA) but conclude that a basic token-based baseline using the words which appear most frequently in the title is the most successful approach at the section level.

Lin et al. [2017] address the sub-article matching problem to "identify all corresponding subarticles in the same language edition". They parse out subarticle candidates, mostly using regular expressions, then use SVMs, Random Forests, Decision Trees, Naïve Bayes, Logistic Regression, and Adaboost to identify subarticles, with Linear SVM and Random Forest being the most successful.

According to Ostapuk et al. [2020] a significant number of Wikidata entries has no corresponding article in any language. Assuming that many of these 'orphan entries' are described in existing Wikipedia articles or their sections and subsection, they map orphans to (sub)sections using graphs and token-key comparison.

# Chapter 3

# Methodology

## 3.1 Structuring Wikipedia

### 3.1.1 Articles and Revisions

Wikipedia keeps track of all revisions resulting from changes made by its editors, with the exception of individual revisions being removed due to issues such as copyright infringement, offensive content, and vandalism [Wikipedia contributors, 2022d]. Each revision has a designated, unique revision ID and can be accessed both through Wikipedia's API and the website of the related Wikipedia article itself. In addition, Wikipedia provides regular dumps of both the most recent and all revisions of all articles.

For the purpose of this thesis, i.e. the creation of an explorative article corpus, the analysis of extraction heuristics and classifiers, and all evaluations, only the latest revisions as of 1 January 2022 as provided in the respective Wikimedia dump were used, with the exception of first preliminary analysis iterations, which used the revision dump from 1 June 2021.

### 3.1.2 Sections, Subsection, and Subsubsections

As Wikipedia articles grow in length over the course of their revision history, editors add structure to them by dividing them into sections, subsections, and subsubsections. In addition to plain text, sections can contain additional elements, such as diagrams, images, charts, and tables. While individual sections can be navigated from the table of contents, and whereas Wikipedia allows for easy access to images and other content via MediaWiki and Wikidata, Wikipedia does not provide built-in tools to extract individual sections, let alone elements embedded in them. Figure 3.1 shows the table of contents box, the beginning of the history section, and the reference section of the Wikipedia article on *CRISPR* from 6 October 2022.

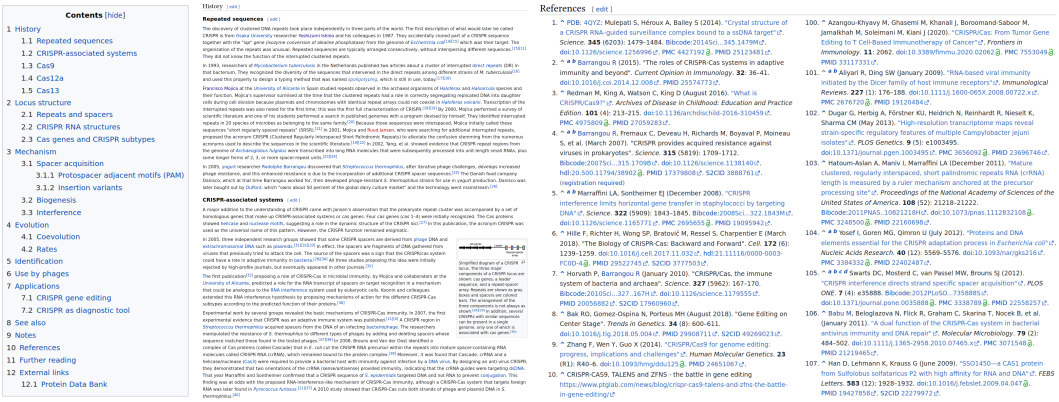


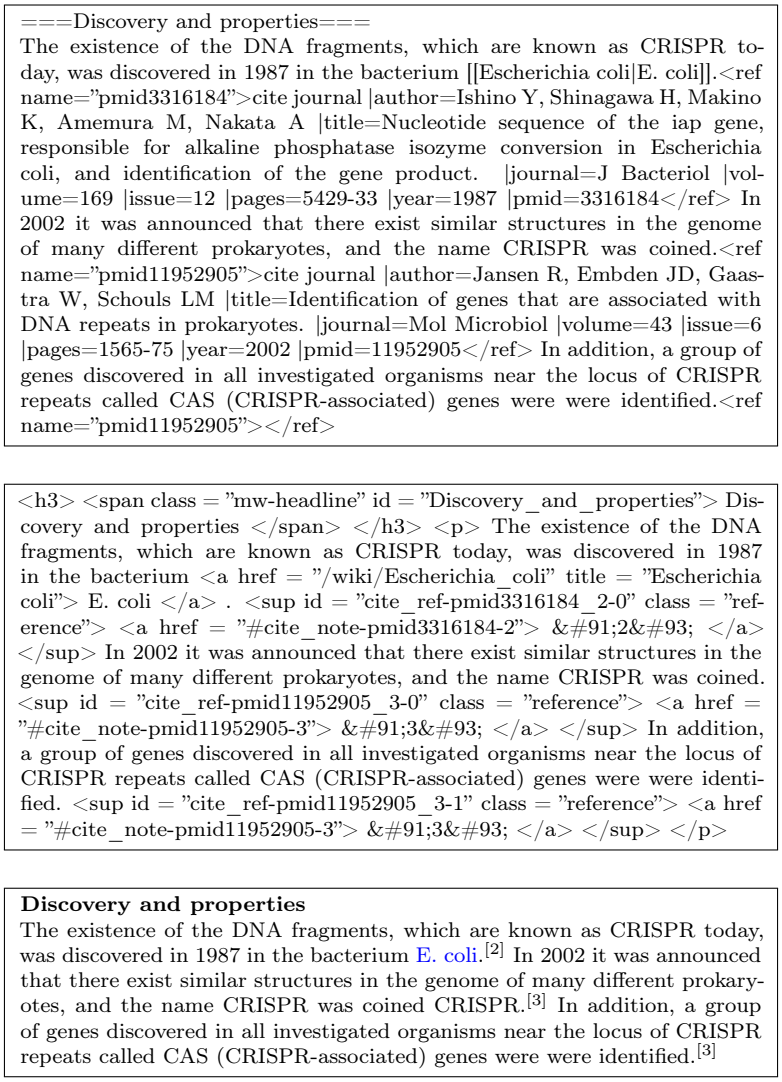
**Figure 3.1:** Contents, Sections and References. [Wikipedia contributors, 2022a]

Owing to Wikipedia's aforementioned guidelines on source accountability, almost all Wikipedia articles have a 'References' and 'Further Reading' as well as other frequent, generic, content-independent sections such as 'External links', 'See also', 'Notes', 'Bibliography', 'Sources', 'Footnotes', 'External sources', or 'Links'. However, there are no guidelines as to which sections a specific type of article should or should not contain, so many section headings are simply editors' choices, resulting in syntactical and semantical variations to represent the section's content [Piccardi et al., 2018]. In contrast, some articles do not contain any sections, with the entire article – with the exception of the introduction and some boilerplate – being made up of a single block of text.

### 3.1.3 Wikidump, API, Browser, Wikitext, and HTML

Articles as provided by Wikimedia dumps and the Wikipedia API on the one hand and the Wikipedia page as displayed in the user's browser on the other might not exactly match up for a number of reasons. The Wikipedia API and Wikimedia dumps store and provide articles in Wikitext, Wikipedia's own markup language, which is parsed to HTML for the purpose of displaying Wikipedia in the browser. While Wikitext provides for powerful templates, updates in template and module syntax can potentially result in faulty output of old revisions, and even at best the "HTML version of a Wikipedia article typically contains more, oftentimes substantially more, information than the original Wikitext source from which the HTML output was produced" [Mitrevski et al., 2020].

Figure 3.2 provides a direct comparison between Wikitext, HTML, and the browser output of a Wikipedia text section. Mitrevski et al. point out that, in order to analyze Wikipedia as seen by the reader, researchers should work



```
===Discovery and properties===
The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium [[Escherichia coli|E. coli]].<ref name="pmid3316184">cite journal |author=Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A |title=Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. |journal=J Bacteriol |volume=169 |issue=12 |pages=5429-33 |year=1987 |pmid=3316184</ref> In 2002 it was announced that there exist similar structures in the genome of many different prokaryotes, and the name CRISPR was coined.<ref name="pmid11952905">cite journal |author=Jansen R, Embden JD, Gaastra W, Schouls LM |title=Identification of genes that are associated with DNA repeats in prokaryotes. |journal=Mol Microbiol |volume=43 |issue=6 |pages=1565-75 |year=2002 |pmid=11952905</ref> In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR repeats called CAS (CRISPR-associated) genes were were identified.<ref name="pmid11952905"></ref>
```

```
<h3> <span class = "mw-headline" id = "Discovery_and_properties"> Discovery and properties </span> </h3> <p> The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium <a href = "/wiki/Escherichia_coli" title = "Escherichia coli"> E. coli </a> . <sup id = "cite_ref-pmid3316184_2-0" class = "reference"> <a href = "#cite_note-pmid3316184-2"> &#91;2&#93; </a> </sup> In 2002 it was announced that there exist similar structures in the genome of many different prokaryotes, and the name CRISPR was coined. <sup id = "cite_ref-pmid11952905_3-0" class = "reference"> <a href = "#cite_note-pmid11952905-3"> &#91;3&#93; </a> </sup> In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR repeats called CAS (CRISPR-associated) genes were were identified. <sup id = "cite_ref-pmid11952905_3-1" class = "reference"> <a href = "#cite_note-pmid11952905-3"> &#91;3&#93; </a> </sup> </p>
```

**Discovery and properties**
The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium E. coli.[2] In 2002 it was announced that there exist similar structures in the genome of many different prokaryotes, and the name CRISPR was coined CRISPR.[3] In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR repeats called CAS (CRISPR-associated) genes were were identified.[3]

**Figure 3.2:** Comparison of Wikitext, HTML, and browser output. [Wikipedia contributors, 2010a,b]

with HTML, but they also note the challenges of processing time (downloading HTML or parsing from Wikitext) and accuracy (arising from template and module version mismatches). For the purpose of this thesis, and to circumvent these restrictions, especially in respect to a prospective analysis of entire revision histories, articles are retrieved from the revision dumps, the Wikitext is extracted for all relevant articles, and Wikitext is processed, cleaned, and handled using a custom parser.

### 3.1.4 WikitextReader

From the more than 16 million entries in the Wikimedia dump, all articles are selected using the Wikipedia namespace for Main/Article (0) [Wikipedia contributors, 2022e], and title, pageid, revid (revision ID), timestamp, and text (i.e. Wikitext) are extracted, yielding a total of 6,129,024[1] articles with an extractable section tree as of 1 January 2022. As Wikitext is partially incomplete and contains in-line references and artifacts, the revisions as provided by the Wikimedia dumps need to be cleaned and processed.

Wikitext does not feature a markup from which a structure can be inferred in the sense that HTML has an inherent document object model tree. A custom WikitextReader therefore processes the Wikitext, extracting headings, associated Wikitext, and categories. Headings are identified by two or more leading '=' at the beginning and the same number of '=' towards the end of the line. '=' characters are cleaned, the line is stripped, comments matching '<.*>' are removed, and categories are identified by lines starting in '[[category:'. The category markup is removed, categories with '|' are cleaned from the character itself and any preceding characters. The section heading tree is built using the length of the '=' heading markup. Based on this heading tree, a section tree can be constructed for each article, featuring the name (heading), level, parent, path, and cleaned text of each section, as well as a list of subsections. The section tree can be recursively parsed for specific headings. The section text pertaining to this heading can then be extracted up to a specific level representing the depth in the section tree.

## 3.2 Finding Science and Technology Articles

### 3.2.1 Heuristic

In order to extract science and technology articles from the revision dump, the articles are filtered using their assigned categories. The main goal here is to find articles that describe innovative technologies as well as scientific concepts, theories, and procedures, i.e. science and technology articles. In addition, only articles with extractable sections are taken into consideration. The corpus does not claim to be complete; instead, it should be as clean as possible in order to be able to find both history and other sections within the same articles, which serve as training data for classifiers and, thus, enable assessment of the most successful strategy/strategies to extract history sections (see section 3.3.3).

[1]The very first corpus extraction iterations were based on the dump from 1 June 2021, which only yielded 6,002,210 articles; see section 3.2.1 for details.

While the list of inclusive strings initially only included the terms 'science' and 'technology' and was later reduced to just 'technolog' (to cover both *technology* and *technological*), the list of exclusive string (hereafter '**stopcats**') was extended over several iterations of manual list expansion and sampling. This approach turned out to be the most viable because, even though Wikipedia's categories span a graph, this "category network is noisy and ill-conceived", as Wikipedia is "notoriously incomplete" [Piccardi et al., 2018] and "authors often tend to overstrain the features" [Schenkel et al., 2007]. In addition, while categories can be very useful in classifying articles, some are simply administrative in nature and only reference the subject matter, whereas others do not imply that the thing described in the article is an instance of the concept the category denotes [Schenkel et al., 2007].

An initial extraction of the June 2021 dump filtering all articles which belong to the categories 'science' or 'technology' reduced the 6,002,210 articles in the dump to 104,155. A small sample of 50 articles included 24 people and 2 companies (48% S&T). When articles belonging to categories containing the strings 'person' or 'company' were excluded, this number decreased to 57,681. A sample of 50 articles from this set contained 13 media entities, 7 people, 6 institutions, and 4 journals (40% S&T). Following a review of the most frequent categories, the stopcat list was expanded to exclude media ('films', 'movies', 'series', 'anime', 'manga', 'books', 'novels', 'screenplays', 'fiction', 'stories', 'games'), people ('people', 'fellows', 'members', 'male', 'female', 'writers', 'authors', 'alumni'), institutions ('institution', 'companies', 'colleges', 'universities', 'council', 'convention'), and some other concepts ('journals', 'magazines', 'births', 'deaths'). This once more roughly halved the number of articles to 27,819. A sample of 50 articles from this set contained 10 institutions, 4 events, and 3 lists (66% S&T). The stopcat list was therefore extended by institutions ('institute', 'department', 'university', 'academies', 'committees', 'commissions', 'foundations', 'associations', 'organizations', 'societies', 'museums', 'establishments', 'agencies', 'churches', 'ministries'), and events ('events', 'festivals', 'conventions', 'awards', 'conferences'), among some other categories, almost doubling them to 55.

Not only did iteration 2, 3, 4, and 5 use the more recent dump from 1 January 2022 onwards, but checks on categories also included the most frequent tokens instead of just the most frequent categories, allowing for easier selection of stopcats. No sample was taken during iteration 2, as some bugs had to be fixed, but the list was extended by a spelling variety of another stopcat ('organisations' in addition to 'organizations'). The sample of iteration 3 contained several glossaries, indices, and authorities, as well as esoteric and pseudoscience articles, so the stopcat list was extended by collections ('indixes', 'indices', 'glossaries'), places ('villages', 'towns', 'cities', 'countries', 'states', 'places'), and articles

**Table 3.1:** Number of science and technology articles in dump, number of categories assigned to science and technology articles, size of sample taken and number of science and technology articles in sample. Preliminary iterations 1.1 to 1.3 performed on dump from 1 June 2021 (6,002,210 articles), improved iterations 2 to 5 performed on dump from 1 January 2022 (6,129,024 articles). No sampling in iteration 2 due to bug fixing. Sample size in iteration 3 and 4 increased to 100 as two samples were taken to assess number of articles with history sections in articles with and without designated history sections. Final iteration with sample statistically significant at CL=95% and CI<5%. A large number of articles has more than one category (c.f. Schenkel et al. [2007]).

| | | Science & Technology | | | |
|---|---|---|---|---|---|
| **Iteration** | **Stopcats** | **Articles** | **Categories** | **Sample Size** | **Positive** |
| 1.1 | 0 | 104,155 | 168,187 | 50 | 24 (48%) |
| 1.2 | 2 | 57,681 | 98,004 | 50 | 20 (40%) |
| 1.3 | 29 | 27,819 | 43,612 | 50 | 33 (66%) |
| 2 | 55 | 17,085 | 18,034 | – | – |
| 3 | 56 | 16,961 | 17,840 | 100 | 88 (88%) |
| 4 | 73 | 15,177 | 14,667 | 100 | 96 (96%) |
| 5 | 79 | 8,402 | 8,752 | 650 | 621 (96%) |

describing unscientific topics ('esoteric', 'pseudoscience'), in addition to some other items. Iteration 4 already scored 96 science and technology articles in the 100 articles sampled. While this ratio was already very satisfying, the sample still contained a large number of articles (*Hydrogen*, *Statistics*, *Geochemistry*, *Political sociology*) and categories ('by country', 'schools', 'occupations', 'parks', 'districts') which proved difficult to assess, so the stopcat list was extended one last time to include certain places ('parks', 'districts', 'by country'), as well as some other categories. Additionally, iteration 5 introduced a second list of stopping strings (hereafter '**stoptitles**'). Article titles are checked for these strings and, if matched, the article in question is excluded. The list contains the strings 'list of', 'index of', ' in ', ' on ', 'history', 'institution', 'company', 'school', 'college', 'institute', 'department', 'university', 'academy', 'committee', 'commission', 'foundation', 'association', 'authority', 'organization', 'society', 'council', 'museum', 'establishment', 'agency', 'church', 'ministry', and 'science and technology'. Table 3.1 shows the results of each iteration, with the evaluation of the final iteration being the first statistical significant one (see Section 4.1.2).

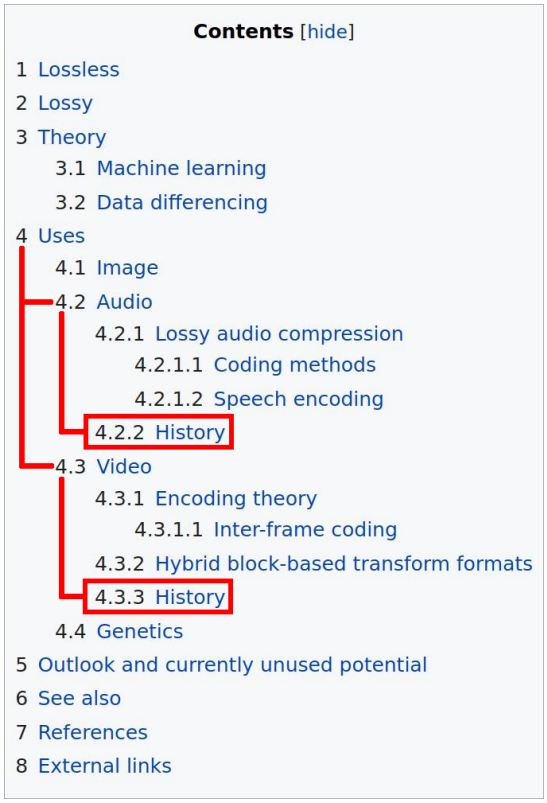



**Figure 3.3:** Two subsubsections in one Wikipedia article describing the history of two application for data compression respectively. [Wikipedia contributors, 2022b]

## 3.3 Finding History Sections

### 3.3.1 Level

Each iteration recorded the number of articles with a section with the heading 'history' (exact match) as well as the number of articles with a section whose heading contained the string 'history' (iteration 1.0 to 2) or either 'history' or 'histori' (iteration 3 to 5) respectively (partial match). One issue that became apparent during iteration 3 was the level at which the history section was located. Figure 3.3 illustrates this: While the Wikipedia article on data compression seems to contain two history sections, both are subsubsections. More importantly even, neither section describes the historical development of data compression, i.e. the improvements in compression ratios since the mid 20th century or a timeline of algorithmic concepts, but the history of the application of data compression in the respective usage scenario (audio and video).

**Table 3.2:** Number of history sections in science and technology articles, both as partial (heading contains string 'history' or 'histori') and exact (heading 'history') match. From iteration 3 onwards the level of the section was recorded.

| | | History Sections in Science & Technology Articles | | | |
|---|---|---|---|---|---|
| | | Heading with 'histor[y\|i]' | | Heading 'history' | |
| **Iter.** | **Articles** | **Any Level** | **Top Level** | **Any Level** | **Top Level** |
| 1.1 | 104,155 | 13,965 (13.41%) | – | 11,145 (10.70%) | – |
| 1.2 | 57,681 | 10,308 (17.87%) | – | 8,066 (13.98%) | – |
| 1.3 | 27,819 | 7,340 (26.38%) | – | 6,288 (22.60%) | – |
| 2 | 17,085 | 4,454 (26.07%) | – | 3,847 (22.52%) | – |
| 3 | 16,961 | 4,743 (27.96%) | 4,564 (26.91%) | 3,953 (23.30%) | 3,861 (22.76%) |
| 4 | 15,177 | 4,093 (26.97%) | 3,933 (25.91%) | 3,419 (22.53%) | 3,332 (21.95%) |
| 5 | 8,402 | 2,363 (28.12%) | 2,289 (27.24%) | 2,068 (24.61%) | 2,021 (24.05%) |

Table 3.2 gives an overview of the number of history sections in each iteration. As, upon closer inspection, many partial-match history sections are not history sections, and because exact-match history sections at any level below the top level only occur in less than one percent of all articles in the corpus, training data for history section classification is sourced from those articles which have a designated section with the exact heading 'history' at top level (hereafter '**designated history section**').

### 3.3.2 Heuristic

The baseline approach to extract history sections is to check all headings and filter out all sections titled 'history'. Sampling during iteration 3 and 4 had shown that, while most sections labeled 'history' do describe the development of the technology featured in the article, a considerable number of articles without designated history sections also have a section which describes its history (hereafter '**non-designated history section**'). For both iterations, 50 articles from both sets (n=100) were randomly sampled and analyzed.

**Table 3.3:** Number of articles with history sections as predicted by article having designated history section as compared to actual number of articles with history section in iteration 3 (left, $N_{total}$ = 1,559, $N_{history}$ = 746, $N_{no\_history}$ = 813, $n_{total}$ = 100, P = 0.98, R = 0.70, $F_1$ = 0.82) and 4 (right, $N_{total}$ = 1,333, $N_{history}$ = 663, $N_{no\_history}$ = 670, $n_{total}$ = 100, P = 0.98, R = 0.72, $F_1$ = 0.83).

| | | **Prediction** | |
|---|---|---|---|
| | Total<br>100 | **History**<br>50 | **¬History**<br>50 |
| **Label** | **History**<br>70 | 49 | 21 |
| | **¬History**<br>30 | 1 | 29 |

| | | **Prediction** | |
|---|---|---|---|
| | Total<br>100 | **History**<br>50 | **¬History**<br>50 |
| **Label** | **History**<br>68 | 49 | 19 |
| | **¬History**<br>32 | 1 | 31 |

Table 3.3 compares the predicted and actual number of articles with history sections. This was more thoroughly assessed during Evaluation I (see Section 4.1.3). Only articles with 10 or more sections (to account for boilerplate sections like 'See also', 'References', 'Further Reading', 'External links', and 'Notes') at top level were taken into consideration in order for the articles to be sufficiently long and have enough exploitable structure (cf. section 3.3.3 and Aprosio and Tonelli [2015]). This yielded 1,559 articles (746 with and 813 without designated history section) and 1,333 articles (663 with and 670 without designated history section) in iteration 3 and 4 respectively. All articles with designated history sections at top level in this corpus happen to have exactly one such section.

### 3.3.3 Classification

Various classifier were trained to find history sections that cannot be extracted by the simple heading-based heuristic described in section 3.3.2. Similar to Aprosio and Tonelli [2015], who use the sections 'Life' and 'Biography' as positive and all remaining sections in the same article as negative examples to identify bibliography sections, articles with designated history sections are extracted and their sections divided into the classes HISTORY and OTHER, which serve as the gold standard data [Field et al., 2020] and, thus, ground truth for training and cross-validation. The first text segments above the list of contents are excluded from training and cross-validation to find the best classifier candidates, as many of them had been labeled as history sections during evaluations (see section 4.1.3). As these section do not have titles, Chen et al. [2021] labeled them 'Introduction'. However, as some articles have designated sections with this title in addition to the first segment, a heading of the format '[ARTICLETITLE] --- Introduction ---' is used here instead.

Articles again need to be sufficiently long and have enough exploitable

**Table 3.4:** Articles and sections (history and other). Boilerplate sections like 'See also', 'References', 'Further Reading', 'External links', and 'Notes' are excluded.

<table>
<tr><td colspan="3">8,402 articles</td></tr>
<tr><td colspan="3">4,409 articles with 3 or more sections (excluding boilerplate)</td></tr>
<tr><td>2,825 without history</td><td colspan="2">1,584 with history</td></tr>
<tr><td rowspan="2">12,520 sections with more than 100 characters</td><td colspan="2">8,179 sections with more than 100 characters</td></tr>
<tr><td>1,574 history sections</td><td>6,605 other sections</td></tr>
</table>

structure to get examples from both classes. The approach used in section 3.3.2 was improved: Instead of only selecting articles with 10 or more sections, an article needs to have three or more sections excluding 'See also', 'References', 'Further Reading', 'External links', and 'Notes'. This increases the number of articles from 1,333 to 4,409, with 2,825 articles without and 1,584 articles with designated history sections. The final training dataset contains 1,574 history and 6,605 other sections, while the 2,825 articles without designated history sections contain a total of 12,520 sections with equal to or more than 100 characters (Table 3.4).

During the cross-validation of Scikit-learn (Sklearn) [Pedregosa et al., 2011] classifiers, two assessments were performed, with the second one excluding sections with less than 100 characters. The second run yielded the same classifier candidates but with two of them having slightly different feature parameters. While the comparison of the two approaches is inconclusive, an oversight during the experiment setup to create the evaluation data to compare the five classifiers in their performance against the articles without designated history sections resulted in sections with less than 100 characters being excluded, the classifiers, however, being set up with the feature parameters according to the candidates resulting from the cross-validation including sections with less than 100 characters. This, however, only affects the Gradient Boosting (4th best) and the Multi-Layer Perceptron (5th best) classifier (see Section 'Sklean & BERT' below for details).

### Sklearn & BERT

26 Sklearn classifiers belonging to different model families underwent a 5-fold-cross-validation to determine the best candidates. Hyperparameters were set to default for all classifiers, with the exception of the four (multi-class) support vector machines, which were also set up with a regularization parameter of 0.025 in addition to the default 1.000, resulting in a total number of 30 classifier setups.

- Nearest Neighbor classifier: K-Nearest Neighbors Classifier
- Naive Bayes classifiers: Bernoulli, Categorical, Complement, Gaussian, and Multinomial Naive Bayes Classifier
- Decision Trees: Decision Tree Classifier, Decision Tree Regressor
- Ensemble classifiers: AdaBoost, Bagging, Random Forest, Extra-Trees, Gradient Boosting, Histogram-based Gradient Boosting
- Neural Network classifier: Multi-Layer Perceptron
- Linear models: Logistic Regression, Passive Aggressive Classifier, Perceptron, Ridge Classifier, Stochastic Gradient Descent, Stochastic Gradient Descent One-Class SVM
- Quadratic Discriminant classifier: Quadratic Discriminant Analysis
- Support Vector Machine classifiers: Linear Kernel SVC, Poly Kernel SVC, Sigmoid Kernel SVC, Radial Basis Function SVC

The training sections are cleaned and preprocessed (lowering, stopping, tokenization). Individual term frequency dictionaries are built for both history and other sections, and the feature vector vocabulary is built from the union of the most frequent tokens in both dictionaries. Several parameters optimize feature selection: oversampling of history sections, unifying year tokens, unifying person tokens, vocabulary sizes of the 10, 100, 1,000, and 10,000 most frequent (non-stopword) tokens, and binary or relative term frequency. The 30 classifier setups thus yield a total of 1,920 classifier-feature setups.

Table 3.5 shows the top results of the cross-validation. Classifier setups with a precision of less than 0.75 or a recall of less than 0.45 are excluded, and results are sorted precision first, recall second. The top five setups are selected, ignoring setups with a classifier ranked higher with different parameters. Table 3.6 shows the ranking when sections with less than 100 characters are excluded.

**Table 3.5:** Classifier performance on training data using 5-fold cross-validation (sections with less than 100 characters **included**); parameters oversampling (OS), mapping of years (Y), mapping of people (P), vocabulary size (V) and term frequency (T); with precision $\geq 0.75$ and recall $\geq 0.45$; sorted by precision first and recall second. Improvement (green), deterioration (red), and agreement (blue), as well as differing feature parameters (underlined) for candidates (gray) over results in Table 3.6.

| Classifier | OS | Y | P | V | T | Precision | Recall |
|---|---|---|---|---|---|---|---|
| Random Forest | 0 | 1 | 0 | 1000 | relative | 0.866 | 0.481 |
| Extra-Trees | 0 | 0 | 0 | 1000 | binary | 0.860 | 0.459 |
| Extra-Trees | 0 | 1 | 0 | 1000 | binary | 0.858 | 0.476 |
| Random Forest | 0 | 1 | 1 | 1000 | relative | 0.856 | 0.498 |
| Extra-Trees | 0 | 0 | 1 | 1000 | binary | 0.855 | 0.457 |
| Extra-Trees | 0 | 1 | 1 | 1000 | binary | 0.844 | 0.456 |
| Extra-Trees | 0 | 1 | 1 | 100 | relative | 0.832 | 0.493 |
| RBF Support Vector | 0 | 0 | 1 | 1000 | binary | 0.832 | 0.482 |
| Extra-Trees | 0 | 1 | 0 | 100 | relative | 0.830 | 0.499 |
| RBF Support Vector | 0 | 1 | 1 | 1000 | binary | 0.829 | 0.487 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Gradient Boosting | 0 | 0 | 1 | 1000 | binary | 0.809 | 0.538 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Multi-Layer Perceptron | 0 | 0 | 0 | 10000 | binary | 0.763 | 0.613 |

Selecting for precision over recall is motivated by the section heading heuristic alone yielding decent (albeit insufficient) results. According to Evaluation I (see Section 4.1.3 for details), roughly 13.24 percent, or 375 of the 2,825 articles without a designated history section should have a non-designated history section. On the other hand, 99.03 percent, or 1,569 ($TP_{history}$) of the 1,584 articles with a designated history section should have a history section, while 15 should not ($FP_{history}$). Given the high precision of the heuristic, using section headings only would result in a precision of around $P_{Heuristic} = 0.990$ and a recall of

$$R_{Heuristic} = \frac{TP_{history}}{TP_{history} + TP_{no\_hist} + FN_{no\_hist}} = \frac{1,569}{1,569 + 375} = 0.807. \tag{3.1}$$

Using the heuristic first and the Random Forest Classifier with a precision of $P_{Classifer} = 0.866$ and a recall of $R_{Classifier} = 0.481$ as fallback would, ideally, increase the overall recall by almost 10 percent while reducing precision loss to less than 2 percent: Given an estimate of 375 articles having a non-designated history section, we would expect a classifier with a recall of 0.481 to correctly

**Table 3.6:** Classifier performance on training data using 5-fold cross-validation (sections with less than 100 characters **excluded**); parameters oversampling (OS), mapping of years (Y), mapping of people (P), vocabulary size (V) and term frequency (T); with precision $\geq 0.75$ and recall $\geq 0.45$; sorted by precision first and recall second. Improvement (green), deterioration (red), and agreement (blue), as well as differing feature parameters (underlined) for candidates (gray) over results in Table 3.5.

| Classifier | OS | Y | P | V | T | Precision | Recall |
|---|---|---|---|---|---|---|---|
| Extra-Trees | 0 | 0 | 0 | 1000 | binary | 0.861 | 0.450 |
| Random Forest | 0 | 1 | 0 | 1000 | relative | 0.857 | 0.482 |
| Extra-Trees | 0 | 1 | 1 | 1000 | binary | 0.855 | 0.461 |
| Random Forest | 0 | 1 | 1 | 1000 | relative | 0.848 | 0.483 |
| Extra-Trees | 0 | 1 | 0 | 1000 | binary | 0.846 | 0.464 |
| Extra-Trees | 0 | 1 | 1 | 100 | relative | 0.845 | 0.495 |
| RBF Support Vector | 0 | 0 | 1 | 1000 | binary | 0.832 | 0.500 |
| RBF Support Vector | 0 | 1 | 1 | 1000 | binary | 0.831 | 0.498 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Gradient Boosting | 0 | 1 | 0 | 1000 | binary | 0.805 | 0.546 |
| Gradient Boosting | 0 | 0 | 0 | 10000 | relative | 0.805 | 0.543 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| Multi-Layer Perceptron | 0 | 1 | 1 | 10000 | binary | 0.761 | 0.598 |
| ... | ... | ... | ... | ... | ... | ... | ... |

identify

$$TP_{no\_hist} = (TP_{no\_hist} + FN_{no\_hist}) \cdot R_{Classifier} = 375 \cdot 0.481 \approx 180 \quad (3.2)$$

articles. Using the above estimate of 180 articles correctly identified by the classifier, we can, given its precision of 0.866, estimate it to incorrectly identify

$$FP_{no\_hist} = TP_{no\_hist} \cdot \frac{1}{P_{Classifier}} - TP_{no\_hist} = 180 \cdot \frac{1}{0.866} - 180 \approx 28 \quad (3.3)$$

articles with a non-designated history section. Under ideal conditions, we can therefore expect that using the heuristic first and the classifier as fallback to achieve an overall precision of

$$P_{Heuristic+Classifier} = \frac{TP}{TP + FP} = \frac{1,569 + 180}{1,569 + 180 + 15 + 28} \approx 0.976 \quad (3.4)$$

and an overall recall of

$$R_{Heuristic+Classifier} = \frac{TP}{TP + FN} = \frac{1,569 + 180}{1,569 + 180 + (375 - 180)} \approx 0.900. \quad (3.5)$$

With this in mind, favoring precision over recall is reasonable with respect to the task at hand: If large spans of the revision history of a Wikipedia article have a (non-designated) history section, it is sufficient to find roughly every other one of them rather than dramatically reduce precision by incorrectly identifying other sections as history.

**Random Forest Classifier**
A Random Forest [Breiman, 2001] is an ensemble classifier which casts a majority vote based on a number of decision trees. It is a combination of tree predictors and works similar to bagging. Each tree is grown from a randomly sampled set of training data and a random selection of features according to which data is split at each node. Random Forest classifiers compare favorably to Adaboost and are very robust to noise, as they always converge and do not suffer from overfitting, even when more trees are added. Despite not being pruned, Random Forests are very fast as the sampled training sets are smaller than the input needed for other decision trees.

**Extra-Trees Classifier**
Extra-Trees ("Extremely Randomized Trees") [Geurts et al., 2006] randomize, either partially or completely, both attribute and cut-point selection while splitting nodes. While the entire learning sample is used, the simplicity of individual node split operations results in Extra-Trees training considerably faster than both Random Forests and Tree Bagging algorithms. A central idea behind Extra-Trees is that randomization and ensemble averaging reduce variance more than other randomization concepts in decision tree algorithms. Error margins are very similar to Random Subspace and Random Forests, and Extra-Trees are as accurate as or even more accurate than other ensemble methods, with Extra-Trees never falling behind on classification problems.

**RBF Support Vector Classifier**
The RBF (Radial Basis Function) Support Vector classifier [Platt, 1999] is a classical support vector machine which tries to fit a hyperplane into the vector space between objects, thereby splitting them into classes while maximizing the gap between them. In cases where the data points are not linearly separable, their vector space is mapped into a higher-dimensional space using a kernel function where they are then separable by a hyperplane. As standard SVMs do not provide posterior probability calibration, an additional sigmoid function is trained to map the output of the SVM to a probability. The parameters of the sigmoid function are fit from the training set using maximum likelihood estimation. As the support vectors of non-linear SVMs often contain a considerable share of the training data, fitting a sigmoid to them can result in bias.

Therefore, hold-out training sets and cross-validation are used as training data for the sigmoid.

**Gradient Boosting Classifier**
Gradient Boosting [Friedman, 2002] finds a classifier $H$ by optimizing a base learner $h$ by iteratively fitting a parameterized function $\beta$ and parameters $a = a_1, a_2, ...$ via additive expansion of the form:

$$H(x) = \sum_{m=0}^{M} \beta_m h(x; a_m) \tag{3.6}$$

Starting from an initial $H_0(x)$, $\beta_m$ and $a_m$ are fit by minimising some loss function $\Psi(y, H(x))$ such as squared-error for $m = 1, 2, ..., M$ steps and N training samples:

$$(\beta_m, a_m) = arg \min_{\beta,a} \sum_{i-1}^{N} \Psi(y_i, H_{m-1}(x_i) + \beta h(x_i; a)) \tag{3.7}$$

with

$$H_{m+1}(x) = H_m(x) + \beta_m h(x; a_m) \tag{3.8}$$

A weak base learner (e.g. a decision tree) is improved by subsampling from the training data at each iteration [Friedman, 2001]. Accuracy and speed of Gradient Boosting can be improved by incorporating randomization, which also guards against overfitting for the base learner.

**Multi-Layer Perceptron Classifier**
A Multi-Layer Perceptron [Hastie et al., 2009] classifier models a combination of inputs derived from features to a (non-linear) target function. Using an activator function $\sigma$, a perceptron derives a vector from the input vector X, which is then mapped to a vector $Y$ of size $K$, with the $k$th value representing the probability for class $k$, with an additional bias in the output layer. An output function $g$ generates the final outputs. In its simplest form in case of linearly separable input vectors, values $x_1$ and $x_2$ (plus a constant $x_0 = 1$) are multiplied with respective weights $w_1$, $w_2$, and $w_0$, and a heaviside function is applied to the scalar product, which maps a given value pair to class 1 or 0 in a single-value vector as output. Training is achieved by adjusting the weights for each misclassified data point by adding the derivative of the data point, multiplied by an initially specified learning rate. For a multi-layer abstraction of this concept, a larger number of hidden layers is chosen, with each layer extracting features from the input layer before it. Multiple layers enable the

separation of non-linearly separable inputs and deriving hierarchical features from the data, and it is generally advisable to prefer too many layers over too few. Model selection is achieved trying a number of random configurations and selecting the one with the lowest error, using the average prediction over a number of independently trained networks, or by using bagging in that average prediction is gained from networks trained from randomly sampled subsets of the training data.

**BERT**
The uncased BERT base model of the Hugging Face Transformers library [Wolf et al., 2020] was compared against the above Scikit-learn classifiers. BERT (Bidirectional Encoder Representations from Transformers) is a multi-layer bidirectional transformer encoder that can create pre-trained deep bidirectional language representations from unlabeled text by analyzing both the left and right context in all layers: "The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications" [Devlin et al., 2018]. Accordingly, only the self-attention head of the BertForSequenceClassification model was trained at a learning rate of 0.001 and using a weight decay of 0.001. Feature selection and optimization is handled by BERT.

Finally, the most promising classifiers were applied to the articles without designated history sections (now including introductory segments), and a statistically significant sample was evaluated by labelers.

# Chapter 4

# Evaluation

Two evaluations assessed the efficacy of the approaches described in Section 3. Evaluation I assesses the heuristic using categories to identify science and technology articles, and it also provides an approximation of the number of articles with history sections, both in the set of articles with and without designated history sections. Evaluation II focuses entirely on classifying history sections, assessing various classifiers in their effectiveness at identifying these sections in articles which do not contain designated history sections.

## 4.1 Evaluation I

### 4.1.1 Setup

From the 8,402 articles in the corpus 4,409 articles with 3 or more sections (excluding 'See also', 'References', 'Further Reading', 'External links', and 'Notes') were extracted, and a sample of 650 articles was drawn: From the 2,825 without a designated history section 340, and from the 1,584 with a designated history section 310 articles were chosen at random, with both samples sizes having a confidence level of 95 and a confidence interval of 5 percent. All 650 articles were labeled by the author with regard to the question whether they belong to the category of science and technology.

The articles were then randomly split into 10 batches, with each batch containing 34 articles without and 31 articles with designated history sections. The batches were distributed among 9 labelers, with one batch being labeled by 5 labelers. For 15 articles, all of them without designated history sections, labelers were unable to determine whether they contained a history section. For two of them in the inter-labeler batch a majority vote decided the label. Labeling was accomplished using a custom tool which provides examples and instructions, and gives users the option to leave articles unlabeled in case they cannot conclusively decide whether or not the article contains a history section.

**Table 4.1:** Inter-labeler agreement for 5 of the 9 labelers involved regarding history section labeling based on one sample batch of 65 articles (Fleiss's Kappa = 0.819).

| **Labelers** | labeler 02 | labeler 04 | labeler 06 | labeler 08 | author |
|---|---|---|---|---|---|
| labeler 02 | - | 0.849 | 0.715 | 0.908 | 0.816 |
| labeler 04 | 0.849 | - | 0.699 | 0.939 | 0.851 |
| labeler 06 | 0.715 | 0.699 | - | 0.753 | 0.752 |
| labeler 08 | 0.908 | 0.939 | 0.753 | - | 0.909 |
| author | 0.816 | 0.851 | 0.752 | 0.909 | - |

Table 4.1 shows the inter-labeler agreement (Cohen's Kappa) for the five labelers of the inter-labeler batch. The agreement between each pair of labelers is strong to almost perfect, with the exception of labeler 06, who still displays moderate to strong agreement when their assessment is compared to those of other labelers. Fleiss's Kappa for the entire group of labelers is 0.819, indicating almost perfect overall agreement.

### 4.1.2 Science and Technology

As the first non-representative assessment during iteration 4 had already suggested, the precision of the heuristic described in Section 3.3.2 is already very high, with more than 95 percent of all articles in the corpus covering topics of science and technology. The number of science and technology articles is proportionally only insignificantly larger in the subsample of articles with designated history sections at about 0.5 percent. Table 4.2 shows the number of science and technology articles in the samples of both subsets as described in Section 4.1.1 of articles as well as in total.

Some articles proved difficult to assess, others were simply incorrect, most of which were retrieved using inconsistent categories. These include business and science parks (*ANAS High Technologies Park*), attractions (*Rush (Thorpe Park)*, *Musical fountain*), government programs (*Green Salt Project*, *BINC*, *Project Galileo*, *Aerospace Cadets of the Philippines*), technology infrastructure stubs (*State data centre*), pseudoscience and science folklore (*UFO conspiracy theories*, *Gremlin*), degrees and certifications (*Master of Business Informatics*, *Certified Forensic Computer Examiner*, *Offensive Security Certified Professional*, *Certified Information Systems Security Professional*, *Apple certification programs*), court cases (*2G spectrum case*), meta articles (*Kazakh Wikipedia*), military equipment and concepts (*Land Warrior*, *Wunderwaffe*), media (*Ghacks*, *PC Perspective*, *Linux Outlaws*, *Film Sack*, *The Naked Scientists*, *Industry Dive*) and lists (*Comparison of crewed space vehicles*).

**Table 4.2:** Number of science and technology articles in corpus as per Evaluation I.

| **Articles** | **Sample Size** | **Science and Technology** |
|---|---|---|
| without designated history sections | 340 | 324 (95.29%) |
| with designated history sections | 310 | 297 (95.81%) |
| Total | 650 | 621 (95.54%) |

At least one article has now been deemed irrelevant to warrant an article of its own and has subsequently been deleted or merged into existing articles (e.g. *Francis Tuttle Technology Center*). The article on *Hohe Salve*, a mountain in Austria ended up in the corpus as it has a transmission mast and has therefore received the category 'Radio technology'. The article *Human capital flight from Iran* describes a phenomenon rather than scientific research or technology.

Figure 4.1 shows the top 20 categories in the corpus, all of which are science and technology, with the majority relating to engineering, physics, chemistry, and life science.

| | | |
|---|---|---|
| **1** | film and video technology | 495 |
| **2** | emerging technologies | 342 |
| **3** | biotechnology | 330 |
| **4** | television technology | 280 |
| **5** | information technology management | 196 |
| **6** | engine technology | 195 |
| **7** | display technology | 173 |
| **8** | nanotechnology | 160 |
| **9** | mobile technology | 144 |
| **10** | gas technologies | 129 |
| **11** | cooling technology | 121 |
| **12** | radio technology | 121 |
| **13** | assistive technology | 118 |
| **14** | automotive technologies | 110 |
| **15** | nuclear technology | 103 |
| **16** | educational technology | 99 |
| **17** | waste treatment technology | 99 |
| **18** | american inventions | 98 |
| **19** | sustainable technologies | 97 |
| **20** | drilling technology | 97 |

**Figure 4.1:** Top 20 categories.

### 4.1.3 History Sections

While the the vast majority of articles with a designated history section do in fact contain a history section, namely 99.03 percent, or 307 out of the 310 articles sampled, the labelers confirmed that a considerable number of articles without designated history sections contain one or more sections that describe the historical development of the technology or concept: According to Evaluation I, 13.24 percent, or 45 out of the 340 articles sampled contain a history section. Table 4.3 provides an overview of the number of history sections in both article sets.

The good true positive and low false positive rates show that having a section with the heading 'history' strongly indicates that a Wikipedia article has a history section and that it is this section which describes the historical development of the featured technology or concept. On the other hand, the high false negative and low true negative rates show that not having a dedicated section titled 'history' does not mean that this article does not contain a history section.

It should also be noted that an article might contain more than one non-designated history sections, either in addition to another non-designated history sections or one section titled 'history'.

Unfortunately, labelers were only asked to determine *if* an article contained a history section, not *which* one it was that made them come to this conclusion. As some of the articles sampled during Evaluation I also were found by the classifiers, these were labeled by the author as part of Evaluation II. About a quarter (7 out of 29 articles) had an introduction which could easily qualify as a history section.

**Table 4.3:** Number of articles with history sections as predicted by article having designated history section as compared to actual number of articles with history section ($N_{total}$ = 4,409, $N_{history}$ = 1,584, $N_{no\_history}$ = 2,825, $n_{total}$ = 650, P = 0.99, R = 0.87, $F_1$ = 0.93).

| | | **Prediction** | |
|---|---|---|---|
| | Total<br>650 | **History**<br>310 | **¬History**<br>340 |
| **Label** | **History**<br>352 | 307 | 45 |
| | **¬History**<br>298 | 3 | 295 |

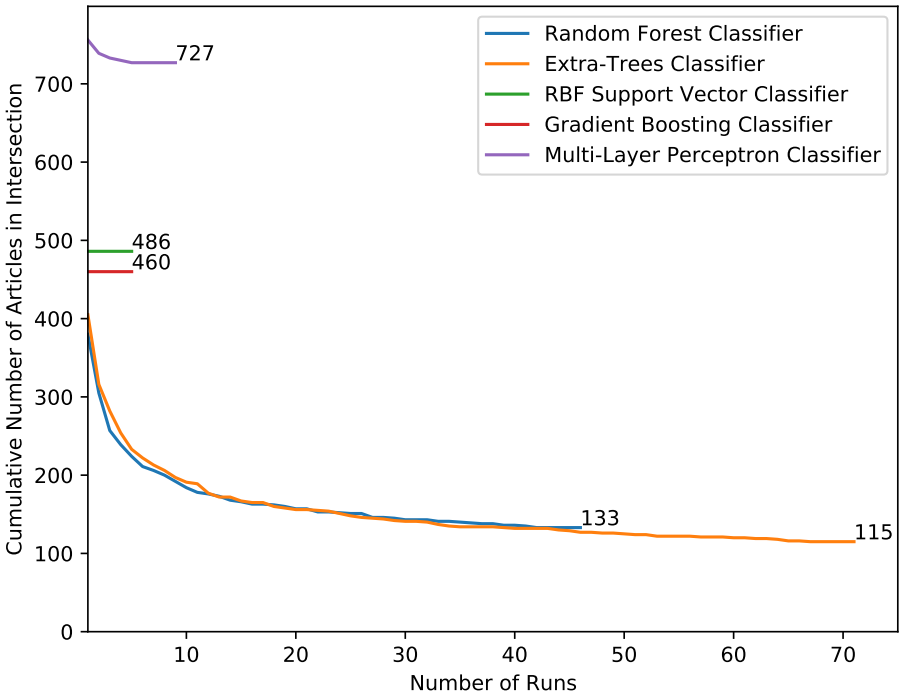



**Figure 4.2:** Number of articles in intersection of runs per Sklearn classifier.

## 4.2 Evaluation II

### 4.2.1 Setup

For Evaluation II the five most promising Sklearn classifiers and BERT were trained using the entire set of 1,584 articles with designated history sections (see Section 3.3.3) and then applied against the 2,825 articles without a designated history section.

As some classifiers are non-deterministic, all Sklearn classifiers were applied against the corpus until the accumulated intersection of articles over all runs did not decrease over five consecutive runs. The Random Forest classifier unifying years, using a vocabulary size of 1,000 and relative token frequency achieved this after 46, the Extra-Trees classifier using a vocabulary size of 1,000 and binary token frequency after 71 runs. The RBF Support Vector classifier and the Gradient Boosting classifier, both unifying persons using named entity recognition, a vocabulary size of 1,000 and binary token frequency basically converged immediately, whereas the Multi-Layer Perceptron classifier using a vocabuly size of 10,000 and binary token frequency required 9 runs. Figure 4.2 compares the convergence curves for all five Sklearn classifiers.

**Table 4.4:** Number of articles assessed to contain history section for each labeler (*new labeler; †93/ ††128 if 29 articles also found in Evaluation I are included).

| Labeler | Articles with History Sections | Articles in Batch |
|---|---|---|
| labeler 00* | 55 | 100 |
| labeler 02 | 62 | 100 |
| labeler 03 | 82 | 100 |
| labeler 04 | 0 (author: 56) | 100 |
| labeler 05 | 39 (author: 75) | 100 |
| labeler 06 | 61 | 100 |
| labeler 08 | 37 | 100 |
| author | 55 | 100 |
| author | 64† | 99†† |
| author | 39 | 85 |

Together with the articles identified by BERT, the evaluation pool contains a total of 1,013 articles. 8 batches of 100 articles were labeled by 8 labelers. No batch was labeled by more than one person; therefore, inter-labeler agreement is not available for Evaluation II, but 7 of the 8 labelers involved had already participated in Evaluation I. However, as one labeler (04) had judged one batch to not contain a single history section, which was in stark contrast to the other batches (see Table 4.4), their judgements were reassessed by the author. One batch of 128 articles contained 29 articles which had already featured in Evaluation I and was labeled by the author, as were 85 articles found by BERT and not included in any of the other batches. In addition, as one labeler (05) was unable to finish the evaluation of their batch, approximately half of that batch was also labeled by the author. A total of 615 articles were labeled as containing a history section. In comparison to Evaluation I, labelers were required to indicate which section or sections it was/were that made them decide that an article contains a history section. Labelers could select more than one section, including the introduction. As only the articles returned by the classifiers were evaluated, the exact number of false negatives for each classifier is unknown.

### 4.2.2 History Sections

Precision and recall are calculated for all classifiers both on article and section level. All sections that were not labeled as history sections by the labelers are considered and therefore labeled as non-history.

For reference, Table 4.5 provides precision and recall for the five Sklearn classifiers as selected during cross-validation and BERT as well as as the

**Table 4.5:** Precion, recall and F-Score of classifiers during cross-validation.

| | **Cross-Validation** | | |
|---|---|---|---|
| | Precision | Recall | $F_{0.\overline{3}}$ |
| Random Forest Classifier | 0.87 | 0.48 | **0.80** |
| Extra-Trees Classifier | 0.86 | 0.46 | **0.79** |
| RBF Support Vector Classifier | 0.83 | 0.48 | **0.78** |
| Gradient Boosting Classifier | 0.81 | 0.54 | **0.77** |
| Multi-Layer Perceptron Classifier | 0.76 | 0.61 | **0.74** |
| BERT | 0.81 | 0.37 | **0.72** |

F-Score, for which $\beta$ is set to $0.\overline{3}$ so that the the models are sorted in the same descending order resulting from the criteria according to which they were selected in Section 3.3.3 (precision $\geq 0.75$, recall $\geq 0.45$, sorted by precision first, recall second).

Table 4.6 shows the results on section level. Precision and recall are calculated for each classifier over the pool of all sections. $\beta$ for the F-Score is again set to $0.\overline{3}$. As only part of the corpus was labeled, namely the articles that were identified to contain history sections by the classifiers, recall is based on the number of 615 articles said to contain history sections as per the labelers' assessments and might therefore be lower than given.

Table 4.7 shows the results on article level. This more lenient approach considers a classifier's decision as correct if it

a) correctly identifies at least one history section (true positive), or

b) ignores the article if it does not contain a history section (true negative)

and only considers the classifier to be wrong if

a) it does not find any history sections even though the article contains one or more history sections (false negative), or

b) none of the sections it identifies are history sections (false positive).

Using this approach, a researcher gets an indication of whether or not an article contains a history section but will have to examine the sections of an article a classifier identified to contain a history section more closely to identify potential false positives or negatives within it.

Results were calculated both including and excluding sections with less than 100 characters but are inconclusive. Removing short sections does not seem to affect the results one way or the other. Tables 4.8 and 4.9 provide results when sections with less than 100 characters are excluded.

**Table 4.6:** Number of history and non-history sections according to classifiers compared to actual number of history and non-history-section. Sections with less than 100 characters **included**.

Random Forest Classifier
P = 0.66, R = 0.17, $F_{0.3} = 0.51$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 305 | **¬History** 8075 |
| **Label** | **History** 1159 | 200 | 959 |
| | **¬History** 7221 | 105 | 7116 |

Extra-Trees Classifier
P = 0.55, R = 0.14, $F_{0.3} = 0.43$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 292 | **¬History** 8088 |
| **Label** | **History** 1159 | 162 | 997 |
| | **¬History** 7221 | 130 | 7091 |

RBF Support Vector Classifier
P = 0.57, R = 0.32, $F_{0.3} = 0.53$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 661 | **¬History** 7719 |
| **Label** | **History** 1159 | 376 | 783 |
| | **¬History** 7221 | 285 | 6936 |

Gradient Boosting Classifier
P = 0.55, R = 0.31, $F_{0.3} = 0.51$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 643 | **¬History** 7737 |
| **Label** | **History** 1159 | 354 | 805 |
| | **¬History** 7221 | 289 | 6932 |

Multi-Layer Perceptron Classifier
P = 0.45, R = 0.45, $F_{0.3} = 0.45$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 1163 | **¬History** 7217 |
| **Label** | **History** 1159 | 526 | 633 |
| | **¬History** 7221 | 637 | 6584 |

BERT
P = 0.50, R = 0.28, $F_{0.3} = 0.47$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 650 | **¬History** 7730 |
| **Label** | **History** 1159 | 328 | 831 |
| | **¬History** 7221 | 322 | 6899 |

**Table 4.7:** Number of articles with and without history sections according to classifiers compared to actual number of articles with and without history sections. Sections with less than 100 characters **included**. Note that labels vary between classifiers depending on each classifier's individual performance.

Random Forest Classifier
P = 0.79, R = 0.17, $F_{0.\overline{3}} = 0.58$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 133 | **¬History** 880 |
| **Label** | **History** 609 | 105 | 504 |
| | **¬History** 404 | 28 | 376 |

Extra-Trees Classifier
P = 0.70, R = 0.13, $F_{0.\overline{3}} = 0.49$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 115 | **¬History** 898 |
| **Label** | **History** 610 | 80 | 530 |
| | **¬History** 403 | 35 | 368 |

RBF Support Vector Classifier
P = 0.60, R = 0.53, $F_{0.\overline{3}} = 0.60$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 486 | **¬History** 527 |
| **Label** | **History** 553 | 294 | 259 |
| | **¬History** 460 | 192 | 268 |

Gradient Boosting Classifier
P = 0.60, R = 0.50, $F_{0.\overline{3}} = 0.59$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 460 | **¬History** 553 |
| **Label** | **History** 559 | 277 | 282 |
| | **¬History** 454 | 183 | 271 |

Multi-Layer Perceptron Classifier
P = 0.53, R = 0.72, $F_{0.\overline{3}} = 0.54$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 727 | **¬History** 286 |
| **Label** | **History** 529 | 383 | 146 |
| | **¬History** 484 | 344 | 140 |

BERT
P = 0.59, R = 0.41, $F_{0.\overline{3}} = 0.56$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 393 | **¬History** 620 |
| **Label** | **History** 565 | 230 | 335 |
| | **¬History** 448 | 163 | 285 |

**Table 4.8:** Number of history and non-history sections according to classifiers compared to actual number of history and non-history-section. Sections with less than 100 characters **excluded**.

Random Forest Classifier
P = 0.66, R = 0.16, $F_{0.3} = 0.51$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 286 | **¬History** 8094 |
| **Label** | **History** 1159 | 189 | 970 |
| | **¬History** 7221 | 97 | 7124 |

Extra-Trees Classifier
P = 0.55, R = 0.14, $F_{0.3} = 0.43$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 300 | **¬History** 8080 |
| **Label** | **History** 1159 | 165 | 994 |
| | **¬History** 7221 | 135 | 7086 |

RBF Support Vector Classifier
P = 0.57, R = 0.33, $F_{0.3} = 0.53$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 665 | **¬History** 7715 |
| **Label** | **History** 1159 | 380 | 779 |
| | **¬History** 7221 | 285 | 6936 |

Gradient Boosting Classifier
P = 0.56, R = 0.31, $F_{0.3} = 0.51$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 645 | **¬History** 7735 |
| **Label** | **History** 1159 | 358 | 801 |
| | **¬History** 7221 | 287 | 6934 |

Multi-Layer Perceptron Classifier
P = 0.44, R = 0.45, $F_{0.3} = 0.44$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 1177 | **¬History** 7203 |
| **Label** | **History** 1159 | 521 | 638 |
| | **¬History** 7221 | 656 | 6565 |

BERT
P = 0.51, R = 0.25, $F_{0.3} = 0.46$

| | | **Prediction** | |
|---|---|---|---|
| | Total 8380 | **History** 572 | **¬History** 7808 |
| **Label** | **History** 1159 | 293 | 866 |
| | **¬History** 7221 | 279 | 6942 |

**Table 4.9:** Number of articles with and without history sections according to classifiers compared to actual number of articles with and without history sections. Sections with less than 100 characters **excluded**. Note that labels vary between classifiers depending on each classifier's individual performance.

Random Forest Classifier
P = 0.82, R = 0.17, $F_{0.3} = 0.59$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 123 | **¬History** 890 |
| **Label** | **History** 611 | 101 | 510 |
| | **¬History** 402 | 22 | 380 |

Extra-Trees Classifier
P = 0.73, R = 0.14, $F_{0.3} = 0.51$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 118 | **¬History** 895 |
| **Label** | **History** 608 | 86 | 522 |
| | **¬History** 405 | 32 | 373 |

RBF Support Vector Classifier
P = 0.61, R = 0.53, $F_{0.3} = 0.60$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 484 | **¬History** 529 |
| **Label** | **History** 550 | 294 | 256 |
| | **¬History** 463 | 190 | 273 |

Gradient Boosting Classifier
P = 0.60, R = 0.50, $F_{0.3} = 0.59$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 468 | **¬History** 545 |
| **Label** | **History** 556 | 279 | 277 |
| | **¬History** 457 | 189 | 268 |

Multi-Layer Perceptron Classifier
P = 0.51, R = 0.72, $F_{0.3} = 0.53$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 730 | **¬History** 283 |
| **Label** | **History** 524 | 375 | 149 |
| | **¬History** 489 | 355 | 134 |

BERT
P = 0.59, R = 0.39, $F_{0.3} = 0.56$

| | | **Prediction** | |
|---|---|---|---|
| | Total 1013 | **History** 369 | **¬History** 644 |
| **Label** | **History** 564 | 218 | 346 |
| | **¬History** 449 | 151 | 298 |

# Chapter 5

# Discussion

## 5.1 Science and Technology

With more than 95 percent of all articles sampled describing science and technology topics, filtering articles by their assigned categories proves successful. Discarding categories iteratively results in a fine-tuned list of excluding categories. However, the resulting corpus is neither perfect nor complete.

On the one hand, some articles missing from the corpus were excluded by their categories. As a very prominent example, the article on the *Large Hadron Collider* belongs to the categories 'Buildings and structures in Ain' and 'Buildings and structures in the canton of Geneva' [Wikipedia contributors, 2022f] and is therefore excluded by the stopcat 'buildings'. Out of 25 technologies to have changed the world according to CNET [Musil, 2020], only 6 are featured in the corpus (*Internet of things*, *Artificial intelligence*, *3D printing*, *Blockchain*, *Quantum computing*, *Drones* [*Unmanned aerial vehicle* in Wikipedia]). While some of the articles in the list, such as *E-cigarettes* (*Electical cigarette* in Wikipedia), might be debatable to begin with, other very valid candidates were excluded as a result of the categories assigned to them. None of the excluded articles had a category that contained the string 'technolog', with the exception of two articles (*Video streaming* [*Video on demand* in Wikipedia] and *Videoconferencing* [*Videotelephony* in Wikipedia]), which, however, contained the stopcat string 'service'. A more severe shortcoming of using Wikipedia's categories to identify articles is revealed by the low number and quality of the categories assigned to some of the other articles in the list missing from the corpus. The article on *Face recognition* (*Face detection* in Wikipedia) belongs to only two categories, namely 'Face recognition' and 'Object recognition and categorization', as does the article on *Autonomous vehicles* (*Vehicular automation* in Wikipedia), with the two categories being 'Uncrewed vehicles' and 'Vehicular automation'.

On the other hand, some articles in the corpus do not really qualify for the concept of science and technology but were included in the corpus based on their categories. As already mentioned in Section 4.1.2, this is again a result of Wikipedia's imperfect category system. To further clean the corpus from articles, additional stopcats could be added. The string 'amusement' would have excluded the article *Rush (Thorpe Park)*. Articles like *UFO conspiracy theories* or *Gremlin* could have been filtered by the strings 'conspiracy' and 'folklore', five of the six media articles (*Ghacks*, *PC Perspective*, *Linux Outlaws*, *Film Sack*, *The Naked Scientists*, *Industry Dive*) by the strings 'website' or 'podcast', and all five degree and certification articles (*Master of Business Informatics*, *Certified Forensic Computer Examiner*, *Offensive Security Certified Professional*, *Certified Information Systems Security Professional*, *Apple certification programs*) by the string 'qualifications'. Other articles could have been excluded by expanding the stoptitle list (*Comparison of crewed space vehicles*). However, some articles will slip past even the most elaborate stopcat list, such as *State data centre*, which has no category disqualifying it from inclusion ('Ministry of Communications and Information Technology (India)', 'Information technology in India', 'E-government in India'), or *ANAS High Technologies Park*, which has but one category ('Science and technology in Azerbaijan'). The same holds for other articles like *Kazakh Wikipedia*, *Green Salt Project*, *BINC*, *Project Galileo*, and *Aerospace Cadets of the Philippines*.

## 5.2 History Sections

### 5.2.1 Heuristic

As expected, designated history sections can reliably be identified by the baseline approach of matching the strings 'history' or 'histori' in their titles. Evaluation I indicates that the heuristic alone is not sufficient though, as there are a considerable number of articles without a designated history section that have a section describing the historical development of a technology. Evaluation II confirmed this assessment, with the number of articles with non-designated history section (615) in the corpus as found by the classifiers and labeled by labelers alone considerably exceeding the extrapolation based on the results of Evaluation I (375).

The most common non-designated history sections labeled as history in the corpus are the introductory section (267), 'Background' (49), 'Development' (41), 'Origins' (31), 'Overview' (16), 'Origin' (8), 'Construction' (7), 'Invention' (5), 'Design and development' (5), 'Discovery' (5), 'Applications' (4), 'Decline' (4), 'Etymology' (4), 'Description' (4), '1980s' (4), '1990s' (4), '2000s' (4), '2010s' (4), '20th century' (4), '21st century' (4), 'Clinical trials' (4), and

'Design' (4). 14 section titles appear only three times, 26 appear only twice, and a staggering 583 headings appear only once. While extending the heuristic by adding common strings like 'background', 'development', or 'origin' would increase the recall of this approach, a considerable number of sections would not match. Conversely, always considering the introductory section as a history section would in turn also increase the number of false positives.

### 5.2.2 Classification

All five Sklearn classifiers and BERT fall behind the expectations based on the cross-validation as outlined in Section 3.3.3.

Excluding sections with less than 100 characters has very little to no effect on recall and precision, with inconsistent improvements not pointing in any clear direction. While excluding sections with less than 100 characters seems to marginally improve precision of the Random Forest and Extra-Trees classifier by 0.03 and RBF Support Vector classifier by 0.01 points respectively on article level, section level performance appears to be unaffected except for the Gradient Boosting classifier, where precision improves by 0.01 points. The Multi-Layer Perceptron classifier seems to fair ever so slightly better when sections with less than 100 characters are included both on article and section level. The precision improves for BERT both on article and section level when short sections are removed, but recall is 0.02 and 0.03 points better when they are included.

The Random Forest classifier scores the best precision, with around a third of all sections identified by it being history sections. However, it also only manages to find less than a fifth of all history sections. The Extra-Trees classifier, the second-best model in the cross-validation, fails completely with the overall lowest recall and yet only a mediocre precision when compared to the other classifiers. The RBF Support Vector classifier achieves the highest F-Score but only manages to identify about a third of all history sections at a low precision below 60 percent. Only the Multi-Layer Perceptron classifier manages to find a satisfying number of articles but scores a precision of below 50 percent. BERT manages to find a quarter of all history sections but labels every other section incorrectly.

The considerable discrepancies between article and section level indicate that classification works well for some articles but completely fails for others. From a pragmatic point of view, using the RBF Support Vector classifier would still return a third of all non-designated history sections. Even though about half of all sections identified are not history sections, a science study researcher would still get an indication of revisions that contain information on the historical development of the technology they are studying and would leave them with less sections to discard than sections gained.

We can therefore reassess the calculations in Section 3.3.3 with the information gleaned from Evaluation II. Using the a section title based heuristic only would give us, again, a precision of $P_{Heuristic} = 0.990$, and a recall of

$$R_{Heuristic} = \frac{TP_{history}}{TP_{history} + TP_{no\_hist} + FN_{no\_hist}} = \frac{1,569}{1,569 + 615} = 0.718. \tag{5.1}$$

Using the heuristic first and the lenient, recall-focused RBF Support Vector classifier with a precision of $P_{Classifer} = 0.61$ and a recall of $R_{Classifier} = 0.53$ as fallback would increase the overall recall by more than 14 percent while reducing precision by almost 9 percent: Given that there are at least 615 articles having a non-designated history section, 294 of which the classifier identifies correctly, and 190 which it identifies incorrectly to contain a history section, using the heuristic first and the classifier as fallback will give us an overall precision of

$$P_{Heuristic+Classifier} = \frac{TP}{TP + FP} = \frac{1,569 + 294}{1,569 + 294 + 15 + 190} \approx 0.901 \tag{5.2}$$

and an overall recall of

$$R_{Heuristic+Classifier} = \frac{TP}{TP + FN} = \frac{1,569 + 294}{1,569 + 294 + (615 - 294)} \approx 0.853. \tag{5.3}$$

Using the heuristic first and the lenient, precision-focused Random Forest classifier with a precision of $P_{Classifer} = 0.82$ and a recall of $R_{Classifier} = 0.17$ as fallback would increase the overall recall by almost 5 percent while reducing precision by less than 2 percent:

$$P_{Heuristic+Classifier} = \frac{TP}{TP + FP} = \frac{1,569 + 101}{1,569 + 101 + 15 + 22} \approx 0.978 \tag{5.4}$$

and an overall recall of

$$R_{Heuristic+Classifier} = \frac{TP}{TP + FN} = \frac{1,569 + 101}{1,569 + 101 + (615 - 101)} \approx 0.765. \tag{5.5}$$

Going back to our initial reasoning, a researcher is, depending on the classifier, provided with a considerable number of revisions with at least one history sections to fill an otherwise lengthy gap in a timeline of revisions without any apparent history section.

# Chapter 6

# Conclusion

This thesis explores the creation and analysis of an explorative corpus of science and technology Wikipedia articles using Wikipedia's category network and classifiers with the aim of assisting science studies research by unlocking Wikipedia's unique position as a community-driven, up-to-date and traceable account of science priority debates.

Using Wikipedia's category system yields a satisfying dataset covering a wide range of articles on innovative technology. While the corpus is not complete in the sense that it does by no means contain all science and technology articles available on Wikipedia, evaluations indicate a very low number of articles which do not fall into this category or are, at most, edge cases. This corpus then serves as the basis for an in-depth analysis of various classifiers regarding their capability for identifying non-designated history sections.

Classification of non-designated history sections is attempted by using sections in articles with designated history sections as positive and negative ground truth and training various Sklearn classifiers and BERT. The performance of all classifiers falls behind the expectations set by the cross-validation with both precision and recall below the thresholds set for the classifier candidate selection. Employing the classifiers as a fallback option in addition to the baseline of using the headings of sections to identify history sections can still aid researchers as it yields more relevant revisions to consider than it leaves irrelevant revisions to discard.

# Chapter 7

# Future Work

In order to increase the size of the Wikipedia science and technology corpus and, therefore, the study of history section mining, all articles in the corpus could be labeled with regard to them covering science and technology topics. This annotation could then be utilized to train classifiers to find additional science and technology technology articles that slipped past the category-based heuristic. This will, however, require identifying and labeling a representative dataset of articles that are *not* science and technology to serve as negative training examples. Alternatively, specific segments other than the categories could be used to classify articles, such as headings, infoboxes, or the introductory section.

While, according to Evaluation I, the assumption that designated history sections are history sections and any other sections are not might be correct, the few incorrectly auto-labeled sections might introduce noise into the training data, which could in turn account for the suboptimal final results. This issue could be addressed by annotating any and all sections in the corpus, including all sections in articles with designated history sections with regard to them being history sections.

Working on section rather than article level could be advisable in general but might cause more harm than it relieves as taking sections out of context can make the already challenging task of deciding whether or not a specific section covers the historical development of a technology even more difficult.

One major pathway to improve history sections classification will be to improve feature selection and fine-tune hyper-parameters of classifiers.

Last but not least, the application of the ideas presented in this thesis can be put to the ultimate test by running a case study using the entire revision history of one or more Wikipedia article(s) and exposing the results to the lens of science study research.

# Bibliography

Alessio Palmero Aprosio and Sara Tonelli. Recognizing Biographical Sections in Wikipedia. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 811–816. The Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1095. URL https://doi.org/10.18653/v1/d15-1095.

Philip Ball. The CRISPR Wars, 2021. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00774-1/fulltext. Online; accessed 6 October 2022.

Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini. Societal Controversies in Wikipedia Articles. In Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo, editors, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 193–196. ACM, 2015. doi: 10.1145/2702123.2702436. URL https://doi.org/10.1145/2702123.2702436.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. WikiTableT: A Large-Scale Data-to-Text Dataset for Generating Wikipedia Article Sections. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 193–209. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.17. URL https://doi.org/10.18653/v1/2021.findings-acl.17.

Jon Cohen. The Latest Round in the CRISPR Patent Battle Has an Apparent Victor, but the Fight Continues, 2020. URL https://www.science.org/content/article/latest-round-crispr-patent-battle-has-apparent-victor-fight-continues. Online; accessed 6 October 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah. A Generative Approach to Titling and Clustering Wikipedia Sections. *CoRR*, abs/2005.11216, 2020. URL https://arxiv.org/abs/2005.11216.

Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. WikiWoods: Syntacto-Semantic Annotation for English Wikipedia. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/432.html.

Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

Jerome H. Friedman. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(01)00065-2. URL https://www.sciencedirect.com/science/article/pii/S0167947301000652.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely Randomized Trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. ISBN 9780387848570. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

Eric S. Lander. The Heroes of CRISPR. *Cell*, 164, 2016. doi: 10.1016/j.cell.2015.12.041. URL https://doi.org/10.1016/j.cell.2015.12.041.

Vincent Larivière and Cassidy R. Sugimoto. The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, editors, *Springer Handbook of Science and Technology Indicators*, Springer Handbooks, pages 3–24. Springer, 2019. doi: 10.1007/978-3-030-02511-3_1. URL https://doi.org/10.1007/978-3-030-02511-3_1.

Heidi Ledford. The Unsung Heroes of CRISPR, 2016a. URL https://www.nature.com/articles/535342a. Online; accessed 28 October 2022.

Heidi Ledford. Titanic Clash over CRISPR Patents Turns Ugly, 2016b. URL https://www.nature.com/articles/537460a. Online; accessed 6 October 2022.

Yilun Lin, Bowen Yu, Andrew Hall, and Brent J. Hecht. Problematizing and Addressing the Article-as-Concept Assumption in Wikipedia. In Charlotte P. Lee, Steven E. Poltrock, Louise Barkhuus, Marcos Borges, and Wendy A. Kellogg, editors, *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 2052–2067. ACM, 2017. doi: 10.1145/2998181.2998274. URL https://doi.org/10.1145/2998181.2998274.

Shan Liu and Mizuho Iwaihara. Extracting Representative Phrases from Wikipedia Article Sections. In *15th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2016, Okayama, Japan, June 26-29, 2016*, pages 1–6. IEEE Computer Society, 2016. doi: 10.1109/ICIS.2016.7550850. URL https://doi.org/10.1109/ICIS.2016.7550850.

Blagoj Mitrevski, Tiziano Piccardi, and Robert West. WikiHist.html: English Wikipedia's Full Revision History in HTML Format. In Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles, editors, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 878–884. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7353.

Steven Musil. 25 Technologies That Have Changed The World, 2020. URL https://www.cnet.com/tech/tech-industry/25-technologies-that-have-changed-the-world. Online; accessed 2 January 2023.

Elizabeth M. Nix. Wikipedia: How It Works and How It Can Work for You. *The History Teacher*, 43(2):259–264, 2010. ISSN 00182745. URL http://www.jstor.org/stable/40543291.

NobelPrize.org. The Nobel Prize in Chemistry 2020, 2022. URL https://www.nobelprize.org/prizes/chemistry/2020/summary. Online; accessed 28 October 2022.

Natalia Ostapuk, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. SectionLinks: Mapping Orphan Wikidata Entities onto Wikipedia Sections. In Lucie-Aimée Kaffee, Oana Tifrea-Marciuska, Elena Simperl, and Denny Vrandecic, editors, *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*, volume 2773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020. URL http://ceur-ws.org/Vol-2773/paper-14.pdf.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. Structuring Wikipedia Articles with Section Recommendations. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 665–674. ACM, 2018. doi: 10.1145/3209978.3209984. URL https://doi.org/10.1145/3209978.3209984.

John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. URL https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf.

Ralf Schenkel, Fabian M. Suchanek, and Gjergji Kasneci. YAWN: A Semantically Annotated Wikipedia XML Corpus. In Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, and Christoph Brochhaus, editors, *Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs*

*"Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany*, volume P-103 of *LNI*, pages 277–291. GI, 2007. URL https://dl.gi.de/20.500.12116/31804.

Semantic Scholar. Semantic Scholar Search, 2022. URL https://www.semanticscholar.org/search?q=crispr&sort=relevance. Online; accessed 4 November 2022.

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic Wikipedia. In Les Carr, David De Roure, Arun Iyengar, Carole A. Goble, and Michael Dahlin, editors, *Proceedings of the 15th International Conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 585–594. ACM, 2006. doi: 10.1145/1135777.1135863. URL https://doi.org/10.1145/1135777.1135863.

Wikipedia contributors. CRISPR — Wikipedia, The Free Encyclopedia, 2010a. URL https://en.wikipedia.org/w/index.php?title=CRISPR&oldid=373513385. Online; accessed 11 October 2022.

Wikipedia contributors. CRISPR — Wikipedia, The Free Encyclopedia, 2010b. URL https://en.wikipedia.org/w/api.php?action=query&format=json&prop=revisions&titles=CRISPR&rvslots=*&rvstartid=373513385&rvlimit=1&rvprop=content. Online; accessed 11 October 2022.

Wikipedia contributors. CRISPR — Wikipedia, The Free Encyclopedia, 2022a. URL https://en.wikipedia.org/w/index.php?title=CRISPR&oldid=1114075441. Online; accessed 6 October 2022.

Wikipedia contributors. Data compression — Wikipedia, The Free Encyclopedia, 2022b. URL https://en.wikipedia.org/w/index.php?title=Data_compression&oldid=1112925399. Online; accessed 6 October 2022.

Wikipedia contributors. Wikipedia:Reliable sources — Wikipedia, The Free Encyclopedia, 2022c. URL https://en.wikipedia.org/w/index.php?title=Wikipedia:Reliable_sources&oldid=1114796177#Primary,_secondary,_and_tertiary_sources. Online; accessed 15 November 2022.

Wikipedia contributors. Wikipedia:Revision deletion — Wikipedia, The Free Encyclopedia, 2022d. URL https://en.wikipedia.org/w/index.php?title=Wikipedia:Revision_deletion&oldid=1120143865#Criteria_for_redaction. Online; accessed 15 November 2022.

Wikipedia contributors. Wikipedia:Namespace — Wikipedia, The Free Encyclopedia, 2022e. URL https://en.wikipedia.org/w/index.php?title=Wikipedia:Namespace&oldid=1112342017. Online; accessed 17 November 2022.

Wikipedia contributors. Large Hadron Collider — Wikipedia, The Free Encyclopedia, 2022f. URL https://en.wikipedia.org/w/index.php?title=Large_Hadron_Collider&oldid=1124049653. Online; accessed 22 December 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association, 2008. URL http://www.lrec-conf.org/proceedings/lrec2008/summaries/420.html.

Michel Zitt, Alain Lelu, Martine Cadot, and Guillaume Cabanac. Bibliometric Delineation of Scientific Fields. In Wolfgang Glänzel, Henk F. Moed, Ulrich Schmoch, and Mike Thelwall, editors, *Springer Handbook of Science and Technology Indicators*, Springer Handbooks, pages 25–68. Springer, 2019. doi: 10.1007/978-3-030-02511-3_2. URL https://doi.org/10.1007/978-3-030-02511-3_2.