Note: This paper was compiled for the 41st PolMeth Meeting, 2024. The second part of the paper showcases the new NLP tool and examines the *Legislative Elites' Reactions to Women's Inclusion*.

Introducing embed2discover: A tool for semi-automated, dictionary-based content-analysis

Laurence Brandenberger[(1,2)1], **Oleg Bakhteev**[(1)], **Jorge M. Fernandez**[(3)], **Sophia Schlosser**[(1)], **Luis Salamanca**[(1)]

(1) ETH Zurich, Switzerland, (2) University of Zurich, Switzerland, (3) Spanish National Research Council (CSIC), Spain

Abstract

We introduce `embed2discover`, a new tool for dictionary-based content analysis. The tool combines state-of-the-art machine learning and language model methodologies with manual inputs from coders to ensure high efficiency, replication, and user control over the annotation process. The tool comes with a four-part training setup and a user-friendly interface. Step 1 expands the dictionary using word embeddings. Step 2 clusters sentences, allowing the user to perform a first coarse classification to speed up the learning process of the classification model. Step 3 uses active learning to refine the classification, and step 4 applies the model to classify the full corpus. We detail the inner workings of `embed2discover`, provide a user-friendly introduction, and showcase the tool on text data from the Swiss parliament.

[1]Contact: Laurence Brandenberger, University of Zürich, Institute of Political Science, laurence.brandenberger@ipz.uzh.ch.
The tool is currently under development and can be made available upon request.

1 Introduction

The use of machine learning (ML) and (large) language models for the content analysis of text-based data has grown in popularity (for primer, see Grimmer, Roberts and Stewart, 2021), but researchers are often weary of employing such methods for fear of retrieving unreliable information (Jordan, Paul and Philips, 2023; Chatsiou and Mikhaylov, 2020; Wilkerson and Casas, 2017). In this paper, we introduce `embed2discover`, a tool for dictionary-based, supervised content analysis of (large-scale) text data. Our tool *assists* human coders (henceforth called 'users') in discovering topics and themes in (large) text corpora and classifying texts by combining the use of state-of-the-art methodologies from natural language processing (NLP) with language models and human annotations.

Traditionally, content analysis is performed as an expert-guided, human annotation process, where researchers devise (elaborate) coding schemes and then proceed to process text data manually and code the read text according to the schemes.[2] One of the most prominent examples refers to the Comparative Manifestos Project (CMP) (Budge, 2001; Klingemann, 2006), where human experts code party policy positions in manifesto texts. Human annotations have also been used to compile datasets. For instance, Nussio and Clayton (2024) follow best practices and hand-code over 80,000 news articles to compile new dataset on lynching events in South America. Other examples include the detection of weak and strong populist elements in transcribed political party broadcasts (Jagers and Walgrave, 2007), or studying polarization dynamics in US congressional hearings on climate change bills using hand-coded discourse network data (Fisher, Waggle and Leifeld, 2013).

To speed up the hand-annotation process, computer-assisted content analysis was developed. Early approaches focus on automated content classifications (e.g., Andersen et al., 1992; Carley, 1994; Cowie and Lehnert, 1996), computer-assisted identification of grammatical patterns (e.g., Franzosi, De Fazio and Vicari, 2012), or topic extraction (e.g., Lee and Kim, 2008). The promise of computer-assisted content analysis is increased efficacy, allowing researchers to either broaden or deepen their analysis through the use of expanding data sources (Grimmer and Stewart, 2013). With recent innovations in natural language processing (NLP) and the advances in (large) language models, computer-assisted content analysis has reached new realms of possibilities (for overviews, see Laurer et al., 2024; Chatsiou and Mikhaylov, 2020). But fears

[2]Note that human annotation has been criticized as well in the literature, especially when it comes to labeling human annotations as 'gold standards' and 'ground truth datasets' (e.g., Song et al., 2020; Mikhaylov, Laver and Benoit, 2012).

potential users have to employ these techniques revolve around replicability, validity, and reliability of the coded results (Jordan, Paul and Philips, 2023; Baden et al., 2022; Muddiman, McGregor and Stroud, 2019).

The concern with using automated content analysis is that both supervised and unsupervised methods usually classify text into predefined categories, either using a dictionary (common in unsupervised approaches) or hand-annotated texts (i.e., sentences, paragraphs, or documents) (Wilkerson and Casas, 2017). Whereas dictionary approaches have considerably sped up the annotation process, they are also heavily biased (Carley, 1990; Vourvachis and Woodward, 2015; Van Atteveldt, Van der Velden and Boukes, 2021). The biggest issue resides in the fact that dictionaries are fixed words (or n-grams) that do not account for (i) linguistic flexibility, (ii) linguistic changes over time and (iii) translation biases (Van Atteveldt, Van der Velden and Boukes, 2021). For supervised methods, the user has to do is to set up a classification model and feed it with hand-annotated texts and allow a model to learn distinguishing characteristics from the text (i.e., existence of words, n-grams, linguistic structures). Then, the supervised machine learning models generally assign weights to these distinguishing characteristics and given enough training data, are able to assign categories to new texts based on the content and the learned weights (for applications, see Hanna, 2013; King, Pan and Roberts, 2013).

There are several drawbacks that currently make researchers weary of applying these unsupervised and supervised models to classify text:

1. Coding schemes based on dictionaries limit the coded texts linguistically, do not account for word changes over time (if temporal data is used, see (Greene, Park and Colaresi, 2019)), and restrict the found texts. This is particularly problematic for concepts that are fussy in nature or have ill-defined boundaries, such as populism, inequality, or biodiversity.

2. For supervised approaches, the researcher does not know apriori how many labeled texts it has to provide the SML in order to achieve a high enough classification score. This makes the use of SML methods less desirable, as it strengthens the idea that these methods are unreliable and constitute a 'black box'.

3. For supervised approaches, the researcher also has to define a clear coding scheme in order to provide the labeled texts. This entails a lot of work.

`embed2discover` forgoes these problems and combines dictionary-based approaches to content analysis with supervised methods. In line with the conclusion from Nelson et al. (2021)

> "these new computer-assisted methods can effectively **complement** traditional human approaches to coding complex and multifaceted concepts in the specialized domain of sociology (and related disciplines), but the evidence is mixed as to whether they can fully replace traditional approaches." (p. 227, emphasis added),

our tool is designed to combine the best of both words: complementing advanced NLP methods with efficient human annotations. The goal of the tool is to help a user hand-label meaningful sentences in text data using a user-friendly UI and train a model to increasingly identify relevant sentences (embedded in paragraphs) by itself. First, the user defines a set of words pertaining to the topic to be discovered. Second, the tool allows the user to train a classification model using active learning. Active learning is powerful as it allows the user control over the annotation process (for a political science application, see Dai and Kustov, 2022). It is set up so that after every annotation step (e.g., when the user has annotated 10 sentences), the classification model is updated to learn from the new information it gains from the newly annotated sentences.

The tool collapses the complexity behind classifications of (large-scale) text data into four distinct steps. **Step 1** entails the expansion of a user-defined dictionary. Here, the user specifies a set of words (min. 5) that pertain to the category of interest. Next, the user specifies the breadth of the dictionary expansion with a set of parameters. Using text similarities based on embeddings, `embed2discover` then proposes a set of closely related words to expand the dictionary. This is particularly useful in cases where the topical category is ill-defined or has fuzzy boundaries. For instance, if a user is interested in identifying biodiversity-related bills in a corpus of legislative bills, then the word 'biodiversity' can be specified in the dictionary (along with a minimum of four other words), and the model then finds closely related words or n-grams, such as 'wildlife' or 'ecological area'. In **Step 2**, `embed2discover` takes the expanded dictionary and selects sentences containing one (or more) of the words or n-grams. The tool then automatically clusters these sentences based on semantic similarities. By clustering similar sentences, the user can then deal with them together, rather than coding them individually. The user is asked to coarsely classify each cluster as 'strongly relevant', 'vaguely relevant' or 'not relevant'. This coarse classification step speeds up the following active learning and allows the researcher to weed out sentences with dictionary words that are not on topic. This is particularly useful for languages where words have increased double meanings and only one meaning pertains to the topic under discovery. **Step 3** is the active learning step where the user refines the classifi-

cation. Here, the user annotates a set of sentences. The user is shown sentences embedded in paragraphs. Previous research advocates for coding text paragraphs or sections over sentences alone (Barberá et al., 2021), which is why `embed2discover` presents the sentence in bold, surrounded by up to 2 sentences pre and post (depending on the structure of the document, i.e., whether or not there has been a paragraph near the selected sentence). Once all sentences in a set are annotated, the model is trained anew, including the newly annotated sentences. After every step, the user is shown progress plots and classification performance scores to get a feeling for the annotation progress. Once the user is satisfied with the model's performance, they can move on to step 4. In **Step 4**, the final model labels all sentences in the dictionary. The user can then download the data alongside final accuracy and performance plots and proceed with their analysis.

2 The Mechanics Behind `embed2discover`

Figure 1 gives an overview over `embed2discover`. The tool consists of an input section, a training section, and an output section. The heart of `embed2discover` is the training section, consisting of 4 distinct steps, where human input is used in every step to keep the training on track. `embed2discover` then outputs labels for each sentence in the text corpus.

2.1 The Text Corpus

The text corpus is set up as non-formatted text files in individual files, each with its own ID. The user can arrange their text files to their own liking. For text preprocessing, we employ the `Spacy` library (Honnibal et al., 2020), and use the `cld2` language detector [3] to handle multilingual corpora. The functionality of the toolbox is based on word and sentence embeddings, i.e., vectorized representations of words and sentences sharing the distributional hypothesis property (Sahlgren, 2008): words and sentences frequently used in the same context have close vector representations.

We complement the text with two sets of embeddings: word embeddings for step 1 (dictionary expansion) and sentence embeddings for step 2 (coarse clustering). The toolbox supports various types of *word embeddings*: pre-trained `word2vec` (Mikolov et al., 2013) and `fasttext` embeddings (Bojanowski et al., 2016), as well as averaged word embeddings derived from contextualized sentence embedding models such as `BERT` (Devlin et al., 2018). For the latter, the tool processes the entire corpus to obtain per-word em-

[3] `https://github.com/CLD2Owners/cld2` For the python version of the library we use package from `https://github.com/GregBowyer/cld2-cffi`.

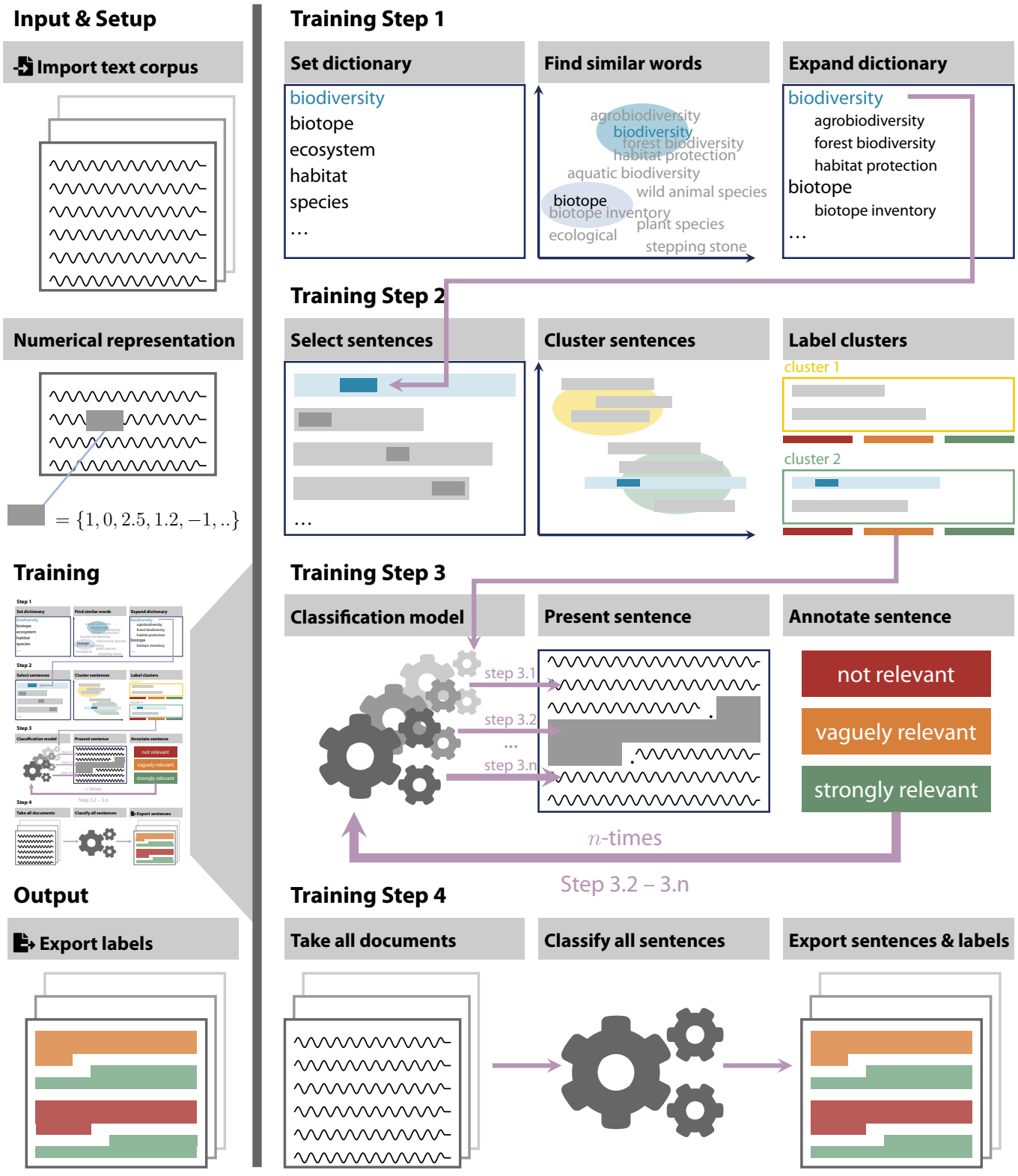


Figure 1: Overview over `embed2discover`.

beddings for each sentence. These word embeddings are heavily dependent on other words presented in the sentences, therefore, we average them across all the sentences and obtain word embeddings with respect to the "average" context of the word in the corpus (Bommasani, Davis and Cardie, 2020). Regarding *sentence embeddings*, the toolbox allows work with models using `Transformers` (Wolf et al., 2019) and `Sentence-BERT` (Reimers and Gurevych, 2019) libraries, including the `SwissBERT` (Vamvas, Graën and Sennrich, 2023) model or the multilingual `Sentence-BERT` models, which allow embedding sentences written in different languages into shared vector space. The tool allows caching embeddings for the target corpus, facilitating the handling of large-scale document corpora without wasting significant computational resources.

2.2 The Four Training Steps

The training is split into four steps (see Figure 1). Step 1 expands the dictionary, step 2 clusters sentences, allowing for a first, coarse classification, step 3 uses active learning to refine the classification and step 4 applies the model to classify the full corpus. For all the steps we calibrate the algorithm hyperparameters (including k in KNN algorithm, classification model parameters, number of clusters in K-means algorithm in step 2) automatically using `Optuna` library (Akiba et al., 2019). The steps 1, 3 and 4 involve a classification model training. For the classification model, we mainly consider kernel logistic regression from (**?**), but any other classification model can be used. We also perform model confidence calibration to make the model confidence aligned with class probabilities. This is important both from the perspective of confidence interpretability and for the active learning step 3, which utilizes the model confidence to obtain new sentences to annotate.

For **step 1**, the user provides a set of keywords (i.e., a dictionary) with words pertaining to the topic to be discovered in the corpus. This dictionary is then expanded. The expansion is based on a k-nearest neighbors (KNN) algorithm and works in two stages. In the first stage, we gather neighbors for each word in the dictionary. In the second stage, we train a binary classification model treating dictionary words and their neighbors as positive class words. For the negative class, we randomly sample words from the corpus. In this setting, we assume that the number of words relevant to the specific task is small in comparison to the full corpus vocabulary, and, therefore, the words randomly sampled from the corpus are unlikely to be relevant dictionary words. Since the next step performance is heavily dependent on the quality of the expanded dictionary we also provide parameters to control the expansion: the model classification proba-

bility threshold, controlling the confidence of the model to add the words into the dictionary, and relative frequency threshold, allowing to discard too frequent words from the dictionary.

Once the dictionary is sufficiently expanded, `embed2discover` moves away from the word level to the sentence level. In **Step 2**, all sentences with the expanded dictionary are identified. We gather all sentences containing at least one word from the expanded dictionary, vectorize them, and perform clustering using K-means algorithm. For the K-means, we use `faiss` library (Johnson, Douze and Jégou, 2019), which provides an efficient implementation of the clustering algorithm for the large-scale problems. Our idea for this step lies in the property of the sentence embeddings: if the sentences with similar meanings lie close enough in the vector space, then during clusterization they will probably be in the same cluster. This step allows us to clusterize them and work with them on a cluster-level instead of considering them individually. For each cluster, a list of sentences is then provided to the user. The user then labels the clusters 'strongly relevant', 'vaguely relevant', or 'not-relevant'. This coarse classification of sentences is used for the first step in the active learning process. In detail, we use a subsample of the sentences in the clustering: the number of sentences to sample depends on the clustering quality and is selected by the user. If the clustering is imperfect, we suggest sampling only a few sentences and improving classification using an active learning procedure. We support two approaches for classification: with respect to class order (i.e., non-relevant < vaguely relevant < strongly relevant) using an ordinal classification approach, and without, treating all the classes as independent and hierarchy-free.

The active learning process (**Step 3**) follows the standard setup: For each round, the user is provided with a set of sentences to label. After completing one labeling round, the classification is updated. By choosing a rather simple model, we can retrain it from scratch after every round. For faster user interaction, we don't perform a hyperparameter selection for the rounds when the number of annotated sentences has not changed significantly. After each round, the user gets active learning progress plots showing the percentage of sentences with matching sentence model prediction and user annotation.

The strategy of sentences to annotate selection can be selected in the configuration file of the tool. By default, we use the following strategy:

1. estimate a best classification model probability threshold by F1-score using a cross-validation procedure;

2. take a pool of sentences with confidence higher than the obtained threshold;

3. from this pool, select sentences with the highest confidence, sentences with the lowest confidence, and sentences randomly sampled from the pool.

The number of sentences in the pool and the ratio of sentences with different levels of confidence can be calibrated in the configuration file. The intuition behind this procedure is the following: we take only sentences in which the model is rather confident, this is obtained using a F_1-score threshold. We mostly obtain sentences randomly sampled from the pool to cover all the confidence levels of the model. We also obtain a subsample of least confident sentences, to cover the borderline cases, when the model is uncertain in its decision, and the most confident sentences, which allows the user to see if the model is overconfident wrongly in some sentences and find out annotation problems in the early stages. For example, this overconfidence can happen when the clustering is performed badly or the annotation is imperfect. Inspired by asymmetric active learning concept (Zhang et al., 2018) by default active learning strategy we don't give the user sentences that the classification model treats as non-relevant or vaguely relevant and process active learning in an imbalanced way. The reason for this is that we are mainly focused on the task of finding sentences that are rare in the corpus, and the ratio of strongly relevant sentences to other sentences is highly imbalanced. However, potentially this can lead to a bad model recall: without providing the classification model with negative examples from the user, the model can overfit the strongly relevant sentences or some specific subsample of such sentences. For this reason, we also support the least confidence active learning strategy that gives the user sentences with the least confidence without thresholding them. The preferred active learning strategy can be set in the configuration of the tool between active learning annotation steps. The tool also supports using multiple active learning strategies in a sequential way.

Once the model is deemed adequate by the user, the user can apply the classification to each sentence in the corpus (**step 4**). At this step, the model is retrained from scratch using the labels from the active learning step. The tool iterates over all the documents from the corpus and applies the classification model to each sentence. The user can select what sentence classes to save during this step. The result of this step is a file with all the classified sentences of the classes required by the user, their context, classification model confidence, and metadata, including the original filename of the sentence, the position of the sentence in the text, and the language of the document.

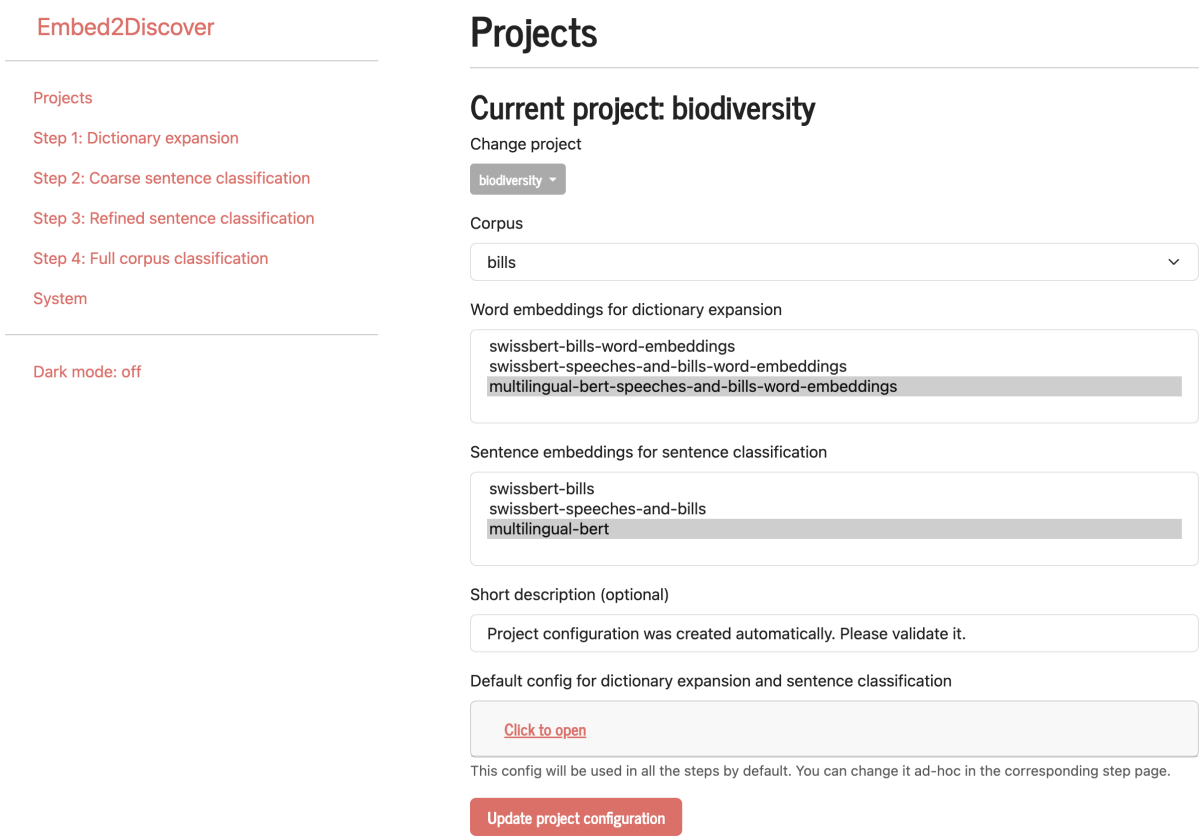


Figure 2: Screenshot of `embed2discover` project set up page.

3 How to Use `embed2discover`

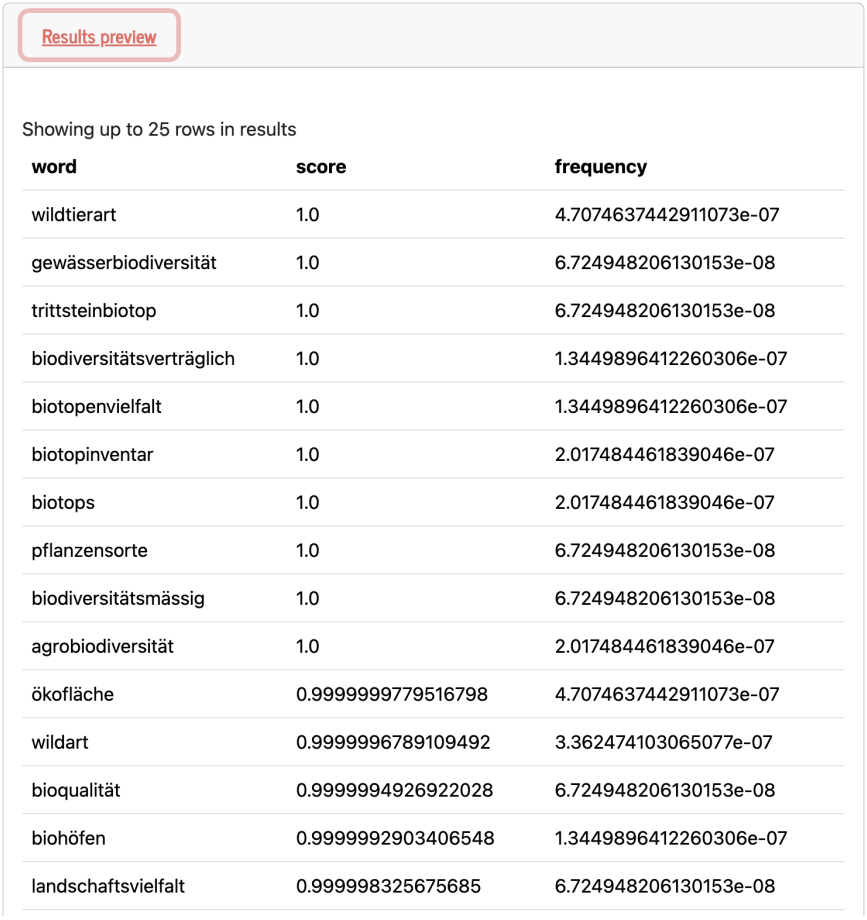
3.1 Input and Setup

Figure 2 shows a screenshot of the `embed2discover` app. The sidebar on the left helps guide the researcher through the different training steps. Project setup entails uploading text (and, if desired, pre-trained embeddings). (Note: The text import button is under development and will be added to the setup page at a later stage.) The user can add a short description, specify the embedding type, and make custom changes in a config file (for advanced users).

3.2 Training Step 1: Dictionary expansion

In step 1, the user specifies the dictionary along with the following settings:

- **Dictionary language**: Set the language of the dictionary.
- **Search in languages**: Texts in which language should be used to expand the dictionary.



Results preview

Showing up to 25 rows in results

word	score	frequency
wildtierart	1.0	4.7074637442911073e-07
gewässerbiodiversität	1.0	6.724948206130153e-08
trittsteinbiotop	1.0	6.724948206130153e-08
biodiversitätsverträglich	1.0	1.3449896412260306e-07
biotopenvielfalt	1.0	1.3449896412260306e-07
biotopinventar	1.0	2.017484461839046e-07
biotops	1.0	2.017484461839046e-07
pflanzensorte	1.0	6.724948206130153e-08
biodiversitätsmässig	1.0	6.724948206130153e-08
agrobiodiversität	1.0	2.017484461839046e-07
ökofläche	0.9999999779516798	4.7074637442911073e-07
wildart	0.9999996789109492	3.362474103065077e-07
bioqualität	0.9999994926922028	6.724948206130153e-08
biohöfen	0.9999992903406548	1.3449896412260306e-07
landschaftsvielfalt	0.999998325675685	6.724948206130153e-08

Figure 3: Screenshot of `embed2discover`: Step 1 results of the dictionary expansion. The first 25 rows in the table are shown to the user for easy validation. Alternatively, the user can download the expanded table and validate it offline.

- **Threshold for classifier confidence [0-1]**: Low values (0.1-.4) expand the dictionary, and higher values restrict the dictionary to the specified words.
- **Expanded dictionary size**: Used to the upper bound of the dictionary.
- **Dictionary**: Text field to specify the dictionary words.
- **Remove frequent words [0-1]**: Specify to remove the top percentage of frequent words in the corpus from the dictionary.

`embed2discover` then performs the expansion and displays the results to the user, see Figure 3.

3.3 Training Step 2: Coarse Classification

In step 2, the user performs a coarse classification with the help of the following settings:

- **Minimal cluster number**: Specify the minimal number of clusters that the k-Means clustering should test for.
- **Maximal cluster number**: Specify the maximal number of clusters the k-Means clustering should test for.
- **(Optional) maximal files to read**: To speed up the process (i.e., for very large corpora), it may be advisable to only read in a portion of the corpus and choose sentences to cluster from the reduced set. The user is advised that the subsequent step (3) still uses the full corpus and is unaffected by the reduced set selected here.

The clustering step currently takes a few hours to run (depending on dictionary size, text complexity, and corpus size). The user is advised with an estimated time-to-completion notification in the app. The app can be closed while the computations are running, and the user can return to t,he app the next day to continue with the manual annotations of the clusters.

Figure 4 shows the UI for coarse classification. For each cluster, the user is presented with up to 20 sentences. The user can go through all clusters and label them. Alternatively, the user can download the clustering data (without sentence count restrictions) and examine them offline. This feature is useful if the cluster size exceeds 100 and the user wants to use search options in the sentences to find relevant clusters.

3.4 Training Step 3: Refined Classification

In step 3, the user activates active learning to refine the classification with the help of the following settings:

- **Language**: Specify the language in which to label sentences. These languages correspond to the ones automatically detected in step 1 upon uploading of the text data.
- **Maximal number of sentences per active learning iteration**: Set the number of sentences the user wants to annotate during one labeling step. To speed up the presentation of new sentences, the user is advised to restrict the number of sentences at this stage.
- **Selecting between using sentences from coarse classification or previous refined classification step**: In the first round, the model pulls sentences from the coarse annotations. In subsequent rounds, the user most likely wants to refine the model (although sometimes the user wants to start over and choose sentences from step 2 in order to restart the annotation process).

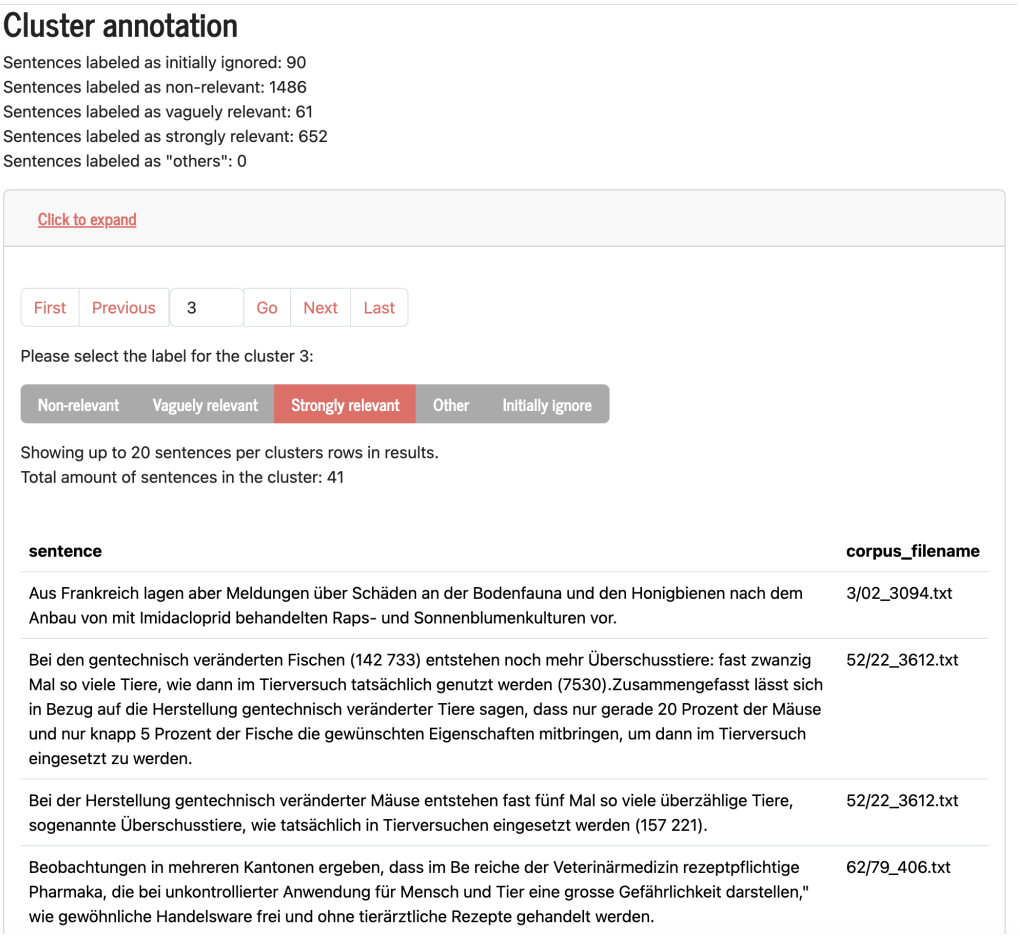


Figure 4: Screenshot of step 2 of `embed2discover`: For each sentence cluster, the user is presented with 20 sentences and can label them as 'strongly relevant', 'vaguely relevant', or 'non-relevant'. If the user wants to ignore the cluster, they can select 'initially ignore' (default value).

Figure 5 shows the UI for the active learning step. The user is informed of the annotation round, how many files have been processed, and how many sentences have been coded so far. Afterward, the user is presented with a set number of sentences for annotation. Upon annotating the sentences, the user can either stop the active learning step (e.g., to take a break or move on to step 4) or continue labeling with a new set of sentences.

Active learning models can be trained indefinitely. The user has to stop the annotation process once the model is deemed good enough. `embed2discover` provides precision-recall as well as additional classification accuracy statistics (f1 and f0.5) together with progress reports to show how the classification improves over time (i.e., after each step).

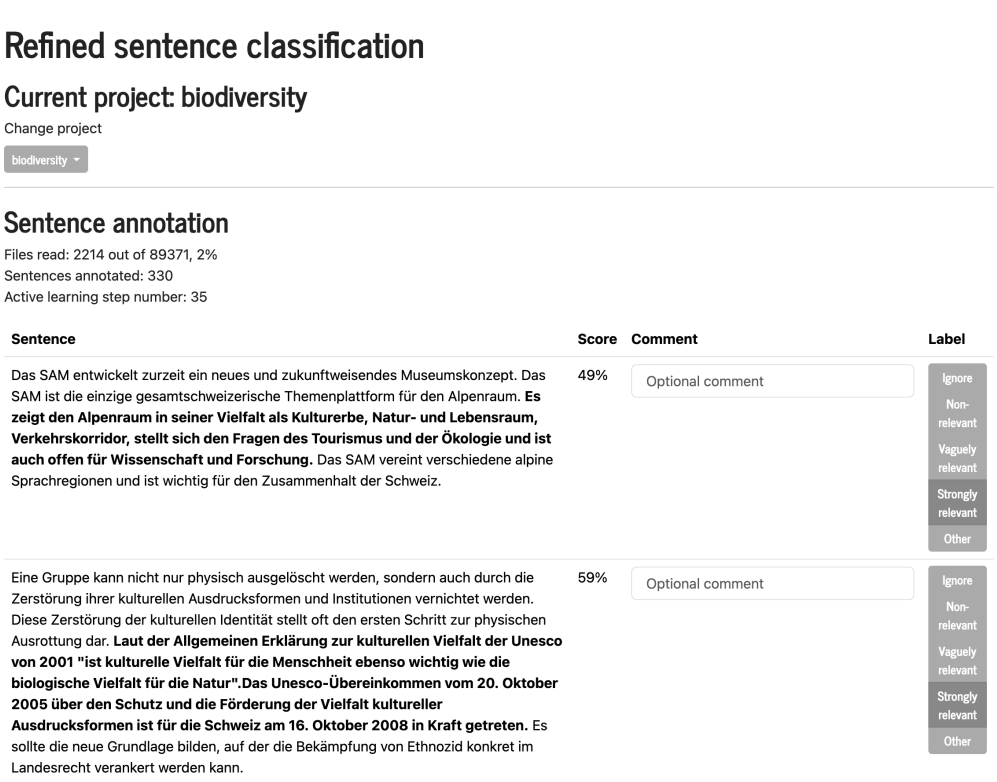
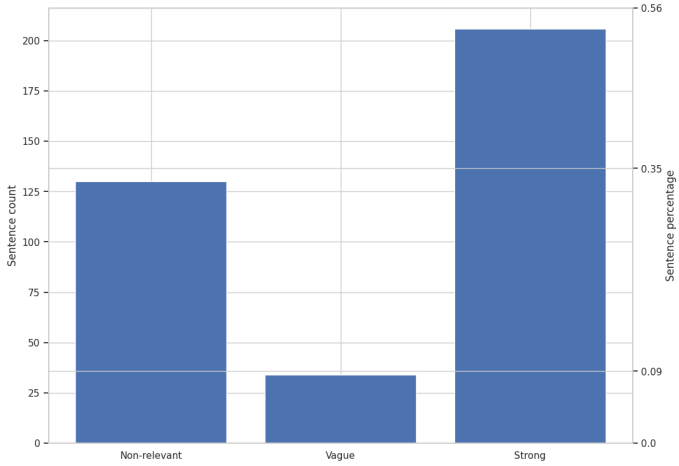


Figure 5: Screenshot of step 3 of `embed2discover`: The user is presented with (embedded) sentences and codes each sentence as ‘strongly relevant’, ‘vaguely relevant’ or ‘non-relevant’.

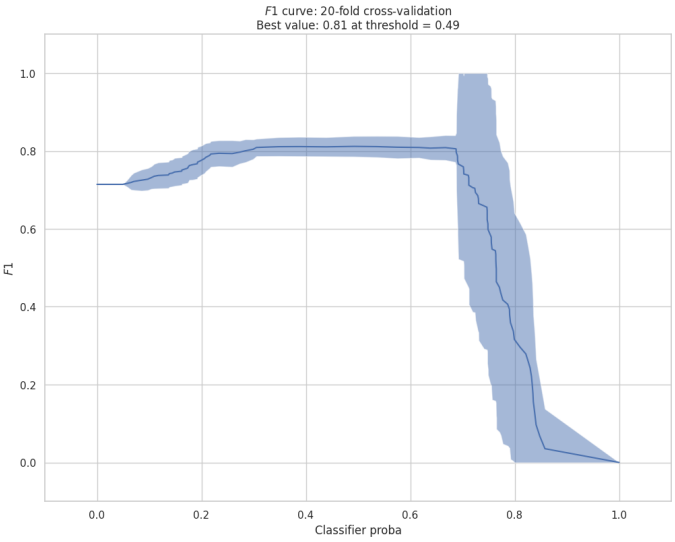
3.5 Training Step 4: Full Corpus Classification

In step 4, the user uses the trained model in order to apply labels to every sentence in the corpus with the help of the following settings:

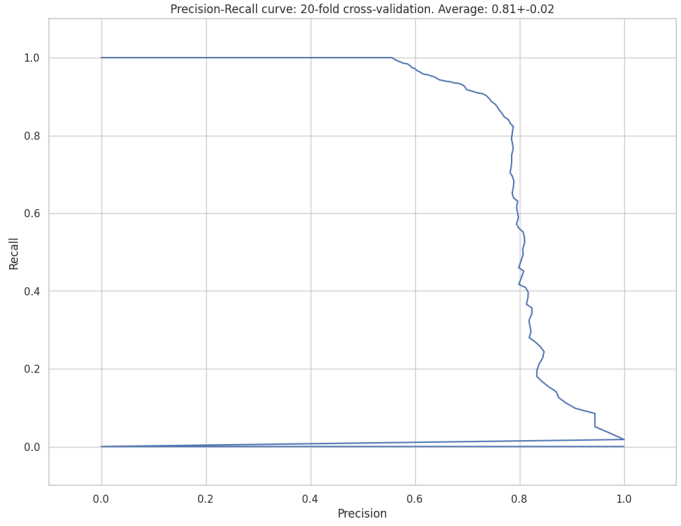
- **Language:** Specify the language in which to label sentences. These languages correspond to the ones automatically detected in step 1 upon uploading the text data.
- **Selecting between using the model from coarse classification or the last refined classification step**
- **Save sentences in the following classes:** The user can save only ‘strongly relevant’, ‘vaguely relevant’, ‘non-relevant’ sentences or a combination thereof.



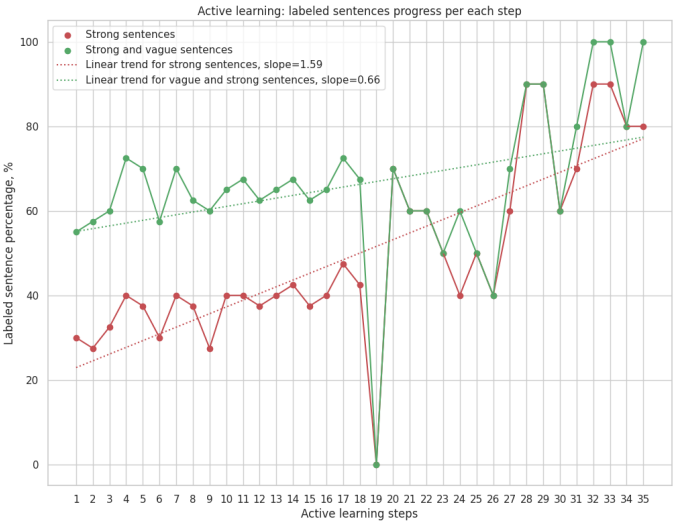
(a) Annotation counts



(b) F1-score



(c) PR-score



(d) Annotation progress

Figure 6: Annotation progress and accuracy plots. These plots are provided to the user after every annotation step in order to advise them about the progress of the annotation.

4 Showcase: Identifying Biodiversity-Related Parliamentary Bills

We showcase the power of `embed2discover` using legislative proposal and bill texts from the Swiss Federal Assembly. The Swiss Federal Assembly was founded in 1848 and has documented its activities since 1891. We have collected data on all legislative proposals and bills (hereafter short: bills) proposed to the Swiss parliament between 1891 and today, including federal reports, enactment drafts, parliamentary and cantonal initiatives, motions, postulates, recommendations (until 2003), interpellations and questions (Salamanca et al., 2024). The resultant dataset consists of over 85k bills, spanning 21 topics (Schlosser et al., 2023).

There are two major challenges when trying to apply dictionary-based classification to this data. First, the data spans over 130 years of legislative activity and reflects significant topical shifts in parliamentary issue engagement over time. Second, not only has issue engagement changed over time, but language itself has evolved over the past 130 years. We use `embed2discover` in order to identify bills related to the overarching topic of *biodiversity*. We have chosen this topic for two reasons: one, biodiversity issues span multiple policy sectors, including the obvious environment, energy, and spatial planning sectors, but also touch upon health, public finances, and international relations. Second, biodiversity issues have risen to the national agenda in the late 1950ies **?**, and before have been addressed under the umbrella of nature conservation and protection (from and of nature). This is ideal for showcasing the power of embedding-based dictionaries and how they can account for language and focus changes over time while still classifying relevant texts.

4.1 Classifying Biodiversity Bills

We start our classification task by setting up a simple dictionary. We focus on words associated with biodiversity, as described in the biodiversity strategy of the Swiss Federal Council[4]. We deliberately chose 'modern' dictionary words to showcase the usefulness of expanding dictionaries using text embeddings. Our dictionary words are: "Biodiversität", "Biodiversitätsprogramm", "biodiversitätsfreundliche", "Waldbiodiversität", "biologisch", "Vielfalt", "Tierpopulation", "Pflanzenpopulation", "Tierart", "Pflanzenart", "Ökosystem", "Biotop", "Lebensraum", "Lebensräume", "Ökosystem", "Ökosystemleistung".

[4] https://www.eda.admin.ch/aboutswitzerland/en/home/umwelt/natur/biodiversitaet.html

We use the default settings in `embed2discover` to expand our dictionary (i.e., classifier confidence = 0.5, expanded dictionary size = 1000, removing 10% most frequent words). Close word matches include: "wildtierart", "gewässerbiodiversität", "biodiversitätsverträglich", "pflanzensorte", "ökofläche", and "landschaftsvielfalt".

In the coarse sentence classification step, we used the default settings (min. cluster size = 3, max. cluster size = 1000, reading in all files). It resulted in 33 clusters (using $N = 2289$ sentences), for which we hand-labeled 652 sentences as 'strongly relevant', 61 as 'vaguely relevant' (i.e., 1 cluster), and 1486 as 'not relevant' (with 1 cluster of 90 sentences set to ignore as it mostly pertained article listings, which are difficult to classify without more context).

In the refined sentence classification step, we used the default settings (10 sentences per run, gathering at most 10 per class) to train our classification model. We trained for 30 steps (i.e., coding 300 sentences), which took approximately 5 hours of annotation time. We stopped the classification once PR hit 80% and of the 300 sentences, more than half were coded as 'strongly relevant'. We have made the experience that the model performs best, once the 'strongly relevant' sentences outweigh the 'not relevant' sentence count. But of course, this depends on the complexity of the texts as well as the complexity of the topic. Figure 7 shows three example sentences from the active learning step with strong relevance.

We then ran the full classification step (step 4) to obtain labels for every sentence in the corpus. Upon exporting, we aggregated the sentences back into our bills dataset, classifying bills as 'biodiversity-bills' once they contained at least 1 sentence marked 'strongly relevant' (at least 3 sentences for bills with more than 100 sentences). This resulted in a total of 4,203 bills (out of 84,918 bills).

4.2 The Benefits of Moving Beyond Dictionary Approaches

We deliberately chose 'modern' dictionary words, i.e., we did not go back into historical texts to see how people have written about biodiversity around this time. The goal is to show that by using embedding similarities, dictionaries can be expanded safely to incorporate language changes over time. However, we look forward to future studies examining dynamic dictionaries and language changes in closer detail. Figure 8 shows word usage in bill texts across time. The term 'biodiversity' was first used in a parliamentary bill in 1991. This is crucial because a simple dictionary-based approach would not find any relevant biodiversity bills prior to 1991 simply because the word was not used. We see, however, that terms such as 'nature' or 'habitat' have been used and correspond quite well to the share of biodiversity bills.

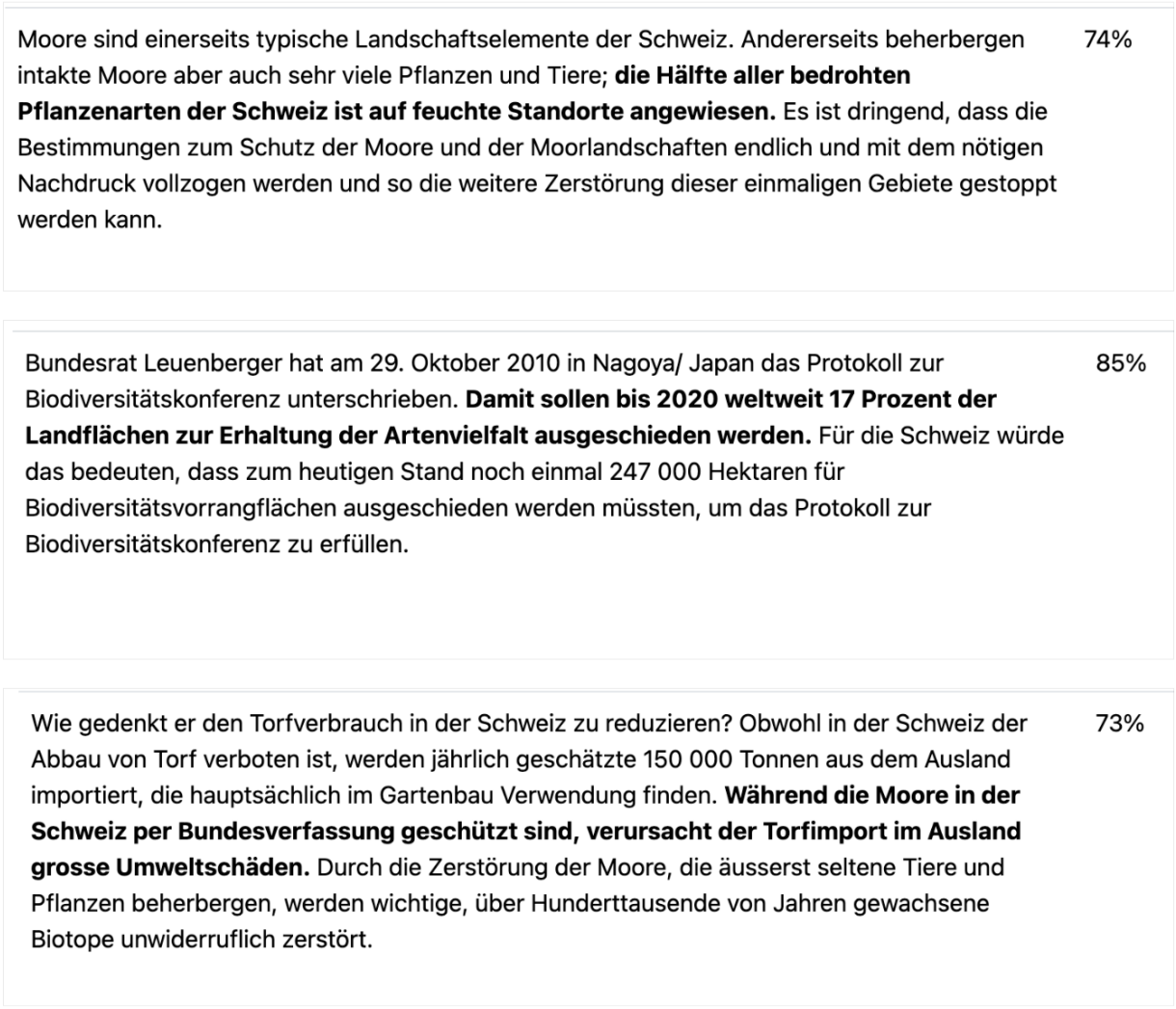


Figure 7: Screenshot of `embed2discover` illustrating three strongly relevant sentences in the refine classification step (step 3).

Taking a closer look at the words used in the biodiversity bills, we see that only a quarter of the bills (27.2%) contain any of the original keywords[5] Figure 9 shows that the keyword 'biodiversity' (fuzzily) matches best. While almost half of the 27.2% bills that use a keyword contain the word 'biodiversity', almost no bills containing the word (0.1%) are left classified as non-relevant. Most bills do not contain any of the 10 dictionary words provided at the start of our classification task. Rather, these bills contain words such as 'nature', 'environment', 'forest', or 'bodies of water'. In Figure 9, we see that the prevalence of these words are similar among both groups (bills with keywords and bills without) and substantively larger than bills classified as 'non-relevant'. `embed2discover` is thus able to expand the dictionary in a meaningful way, allowing for more encompassing topical classifications of texts.

[5]Note that we do not do exact matches. Rather, we use fuzzy matching, allowing word variations (as well as plurals, unfinished words, or wrong spellings).

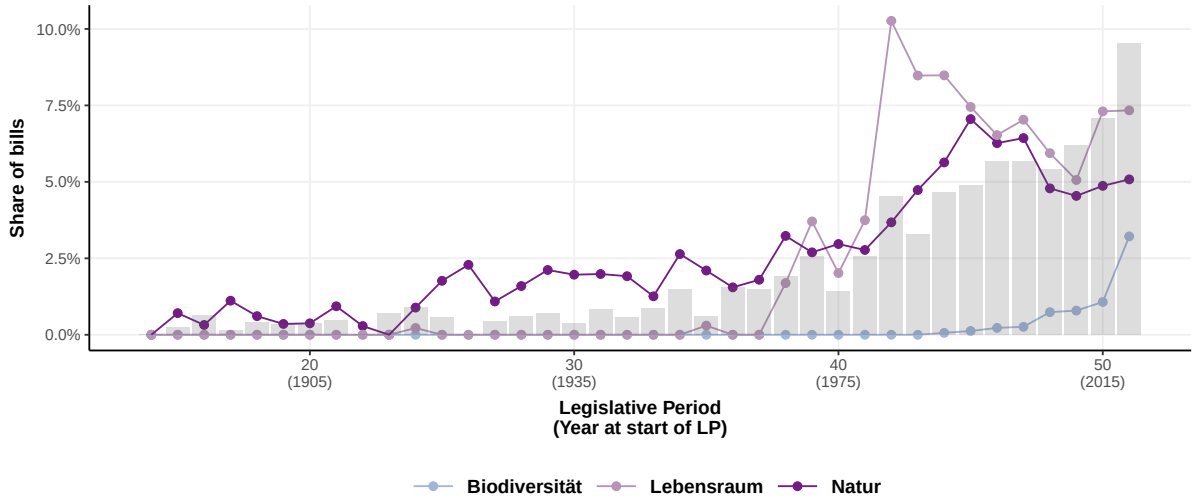


Figure 8: Word usage changes over time. Whereas dictionary approaches fail if words are not present in the texts, `embed2discover` can extrapolate similar words and find close associations.

4.3 Biodiversity Across the Swiss Political Landscape

Figure 10 maps the biodiversity bills into a two-dimensional issue space. Biodiversity bills can be found in every one of the 21 political domains of the Swiss political landscape. Of course, the most predominant domains are Agriculture (N = 1502, 35.7%) and Environment (N = 1001, 23.8%). Biodiversity issues are also addressed in the context of Spatial Planning (N = 306, 7.3%), Energy (N = 247, 5.9%), (Public) Transportation (N = 223, 5.3%) or in International Affairs (N = 147, 3.5%). The bills submitted to the different policy domains cover their unique issues (e.g., Health) while specifically addressing issues of biodiversity, landscape protection, and flora and fauna.

Over the past 130 years, biodiversity issues have gradually gained more space in the federal parliament. Figure 11 shows how biodiversity bills have increased their share from 0.23% in the 15th legislative period to around 9.5% of all bills submitted to the parliament in the 50th legislative period. While researchers argue that Switzerland has ignored biodiversity issues until the late 1950ies (Jaligot et al., 2019), it is clear that some parliamentarians, parliamentary groups as well as the Federal Council have addressed issues of landscape and nature protection. The first biodiversity bill, for instance, deals with the river correction and its impact on the surrounding landscape. In 1925, MP Ryter submitted a bill asking the federal government to take a position on how the new firing range would affect local forests. Interestingly, the vast share of biodiversity bills (87%, compared to the global average of 74%) are submitted as personal bills (parl. initia-

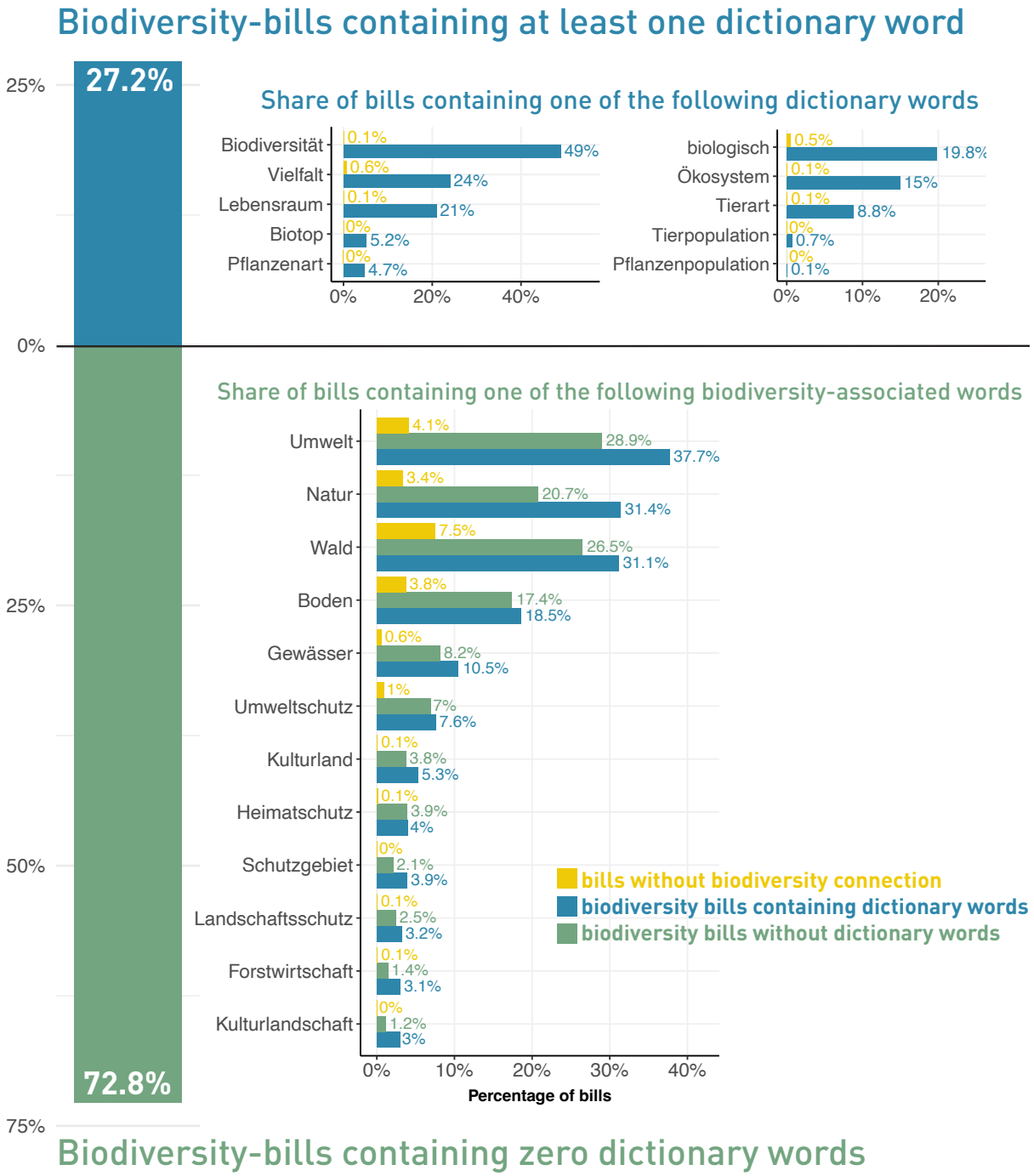


Figure 9: Examination of how dictionary-based approaches fail to discover relevant texts if the dictionary is not expanded using embeddings. Of the 4,203 bills identified as biodiversity-centered, only 27.2% contain words from the dictionary (fuzzy matching enabled).

tives, motions, postulates, interpellations, questions), where MPs act as sponsors of these bills. Historically, MPs from the conservative parties have submitted the largest share of these bills, a trend that has shifted over time. Over time, socialist parties have taken on more biodiversity issues, holding both camps in near balance in the present legislative periods.

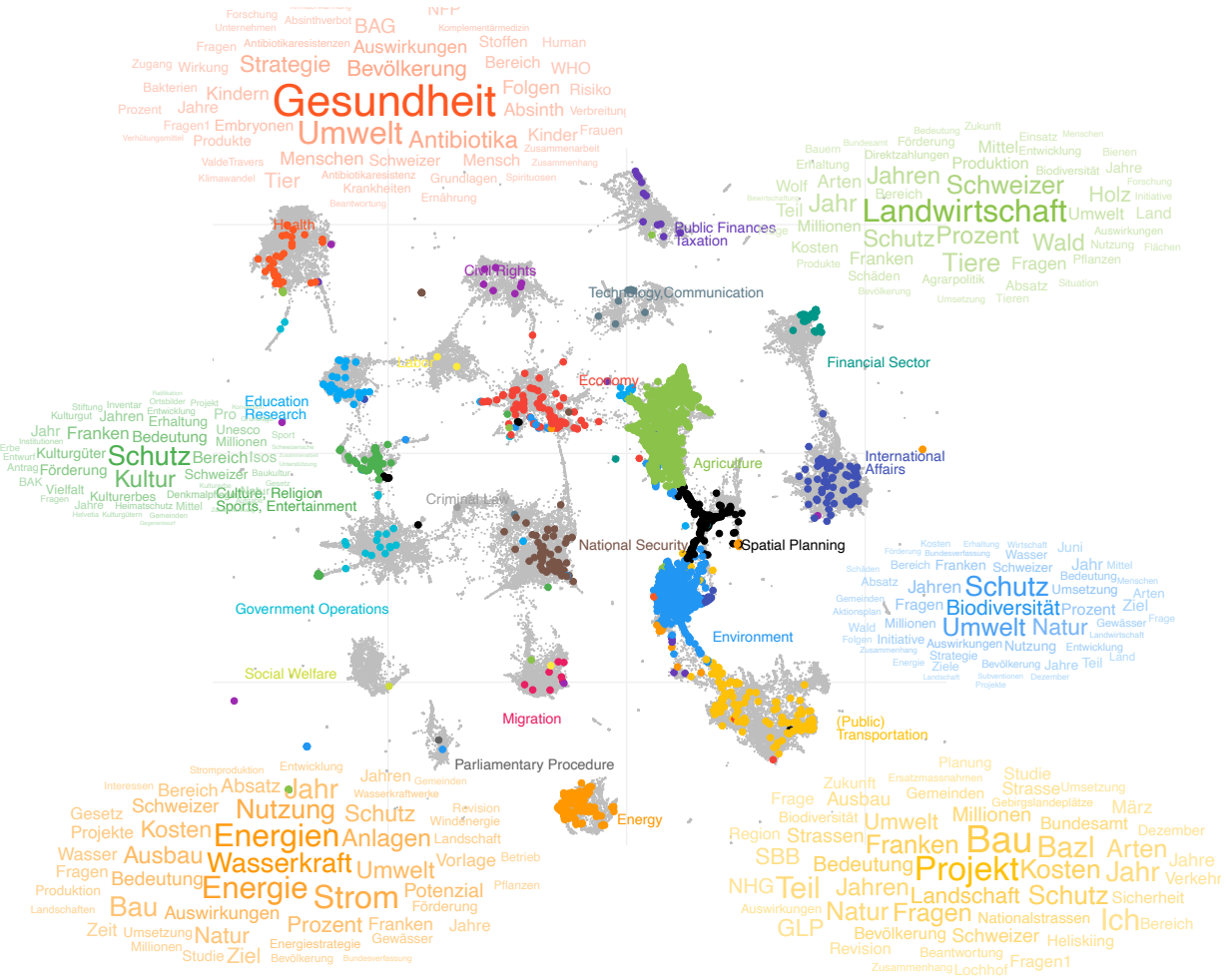


Figure 10: Biodiversity bills (colored dots) are plotted into a two-dimensional issue space. They cover all 21 policy domains, some to a higher degree than others. The word clouds represent the top 50 nouns associated with these biodiversity bills.

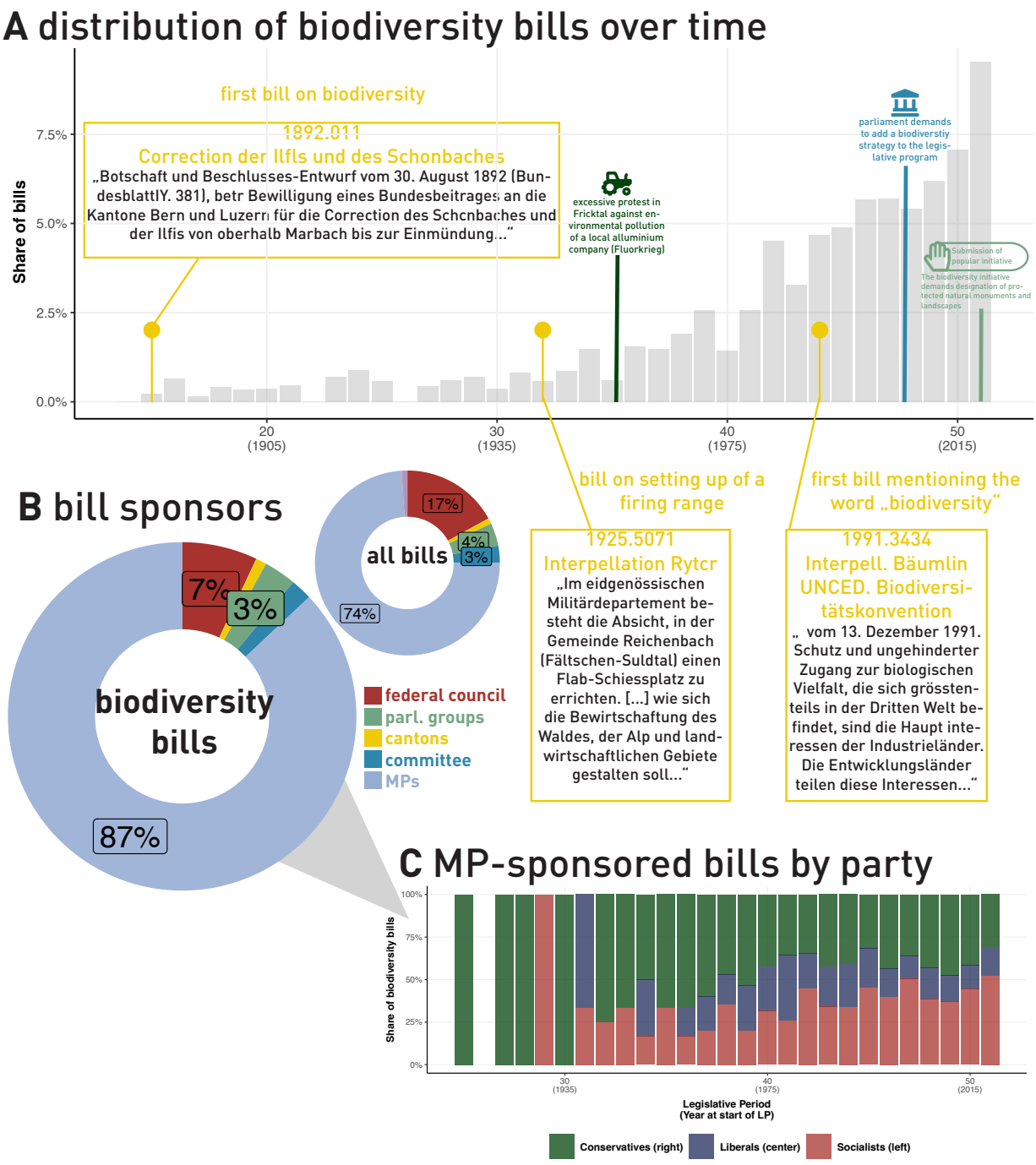


Figure 11: Development of legislative bills on biodiversity issues over time. **Panel A** shows the share of bills submitted to parliament (both chambers) over time, marking key points in Swiss landscape and nature protection efforts. **Panel B** depicts the source of sponsors, with the lion's share of biodiversity bills submitted by MPs themselves. **Panel C** examines MP-sponsored bills in closer detail, depicting the party affiliations of sponsors over time.

5 Showcase 2: After the Enfranchisement: Legislative Elites' Reactions to Women's Inclusion

In our second showcase, we want to understand whether members of parliament update their behavior when their constituency expands. Switzerland provides an interesting case study, as women's suffrage was only introduced in Switzerland in 1971. Despite early attempts (the first one in 1868), it took Swiss politicians and the Swiss government a long time to grant the women to vote and be elected (WR). In 1918, two motions were submitted to the parliament demanding women's right to vote, but even though both passed a parliamentary vote, the motions were ignored by the Federal Council tasked to draft an enactment. In 1951, the Federal Council called the WR 'premature', and only in 1957 did they finally pass a bill to parliament. This bill eventually passed, but as it was subject to the referendum right, Swiss male voters decided against granting WR in 1959 (by popular vote, 33% in favor). In 1968, the Federal Council wanted to sign the human rights convention but was stumped by the fact that Swiss women were not granted political rights. A second campaign was launched, and in February 1971, Swiss male voters granted women the right to vote (for more details on the Swiss women's rights movement, see Figure 18 in the Appendix).

While previous studies on enfranchisement focus primarily on *how* rights were gained (Acemoglu and Robinson, 2000; McConnaughy, 2013; Teele, Kalla and Rosenbluth, 2018), fewer studies focus on how representation styles have changed and how politicians took on their role to represent a broader constituency. Similarly, previous work on enfranchisement has often focused on electoral mobilization and how group identities shift as newly enfranchised join the constituency (Corder and Wolbrecht, 2006; Berlinski, Dewan et al., 2011; Morgan-Collins, 2021; Skorge, 2023). We are interested in the supply side of the enfranchisement dynamics in Switzerland and how political parties and legislators change their legislative behavior to accommodate the newly enfranchised electorate. Therefore, we ask: **Do legislators give more attention to the policy demands of newly enfranchised groups after enfranchisement?**

In order to tackle this research question, we need (i) to know the political demands made by women before the enfranchisement and (ii) to know what MPs have proposed in parliament and whether or not it aligns with the women's demands. Using qualitative analysis based on press releases from the major women's organization in Switzerland (Schweizerischer Verband für Frauenstimmrecht; later renamed to Schweizerischer Verband für Frauenrechte), we have identified four major (Swiss-centric) women's demands:

1. **demand for political rights**: including the right to vote, be elected to cantonal and national parliaments, and equal representation in political offices

2. **demand for preservation and acquisition of citizenship rights upon marriage**: women lost their citizenship if they married a man without Swiss citizenship (the reverse was not the case, i.e., men did not lose Swiss citizenship if they married a foreign woman).

3. **demand for equal opportunities in education and labor**: including equal pay for equal work and no discrimination in education.

4. **demand for the end of insurance discrimination**: including equal insurance premium for men and women.

5.1 Classifying Women-Centric and Women's Demands Bills

In order to identify whether the Swiss parliament tackles women's issues (broad and narrow), we base our analysis on the complete records of bills proposed to the Swiss parliament since 1891. We purposefully use a long time range to describe how the Swiss parliament has tackled women's issues over time, not just those related to women's enfranchisement.

We train five different `embed2discover` models (see Table 1). The first model is a simple model that codes women-centric bills. We start off with nine words indicating 'women' in German. We expand the dictionary to words strongly related to women and then actively train the model on identifying sentences that have something to do with women's issues (including maternity leave, health issues, pay issues, women's immigration issues, violence against women, etc.). Our coding efforts result in a very broad view of women-centric issues. Out of the 85k bills, we assign over 12.5k bills to being women-centric.

The second, third, fourth, and fifth models all train a very specific `embed2discover` model, identifying one of the four women's demands between 1955 and 1975. Table 1 summarizes the model specifications for these four models. The most important change we made was that we narrowed the expansion of the dictionary to a minimum. When working with very narrow topics, we found it most helpful to narrow the dictionary not to include too many adjoining issues.

Table 1: Overview of the five different `embed2discover` models. C1 codes women-centric bills with a simple women-centric dictionary and trained on sentences identifying women's issues (broad coding). C2.1-C2.4 code very narrow women's demands identified from qualitative texts around 1955-1975.

Description	C1 **women-centric bills**	C2.1 **demand for political rights**	C2.2 **demand for preservation and acquisition of citizenship rights upon marriage**	C2.3 **demand for equal opportunities in education and labor**	C2.4 **demand for end of insurance discrimination**
Direct quotes from the Swiss women's organization fighting for women's right to vote and be elected (SVF)		"Vergessen wir aber nicht, dass bis zu dieser Abstimmung 30 Mal in den Kantonen über das Frauenstimmrecht abgestimmt worden war und auch die Erteilung der kleinsten politischen Rechte an die Frau durch mehr als ein halbes Jahrhundert auf starken Widerstand stiess. Die Beharrlichkeit der um ihre Rechte kämpfenden Frauen - ohne in Exzesse zu verfallen - bedeutet eine tiefe Überzeugung, für eine gerechte Sache einzutreten und einen unerschütterlichen Glauben an die Staatsform der Demokratie für alle Erwachsenen." (1971)	"Zur Rechtsungleichheit der verheirateten Schweizerin. [...] Einleitend stellt der Schweizerische Verband für Frauenstimmrecht fest, dass eine offensichtliche und stossende Rechtsungleichheit besteht zwischen dem Schweizer, welcher eine Ausländerin heiratet, und einer Schweizerin, welche sich mit einem Ausländer verheiratet, insofern als erstere übertragen erhält von Gesetzes wegen eine Nationalität auf seine Ehefrau. Der letzteren kann als Folge davon, dass die zuständigen Behörden ihrem Ehemann eine Aufenthaltsbewilligung oder Niederlassung verweigern, so dass sie sich gezwungen sieht, ihren Wohnsitz entweder aus ihrer Heimat oder der Trennung von ihrem Ehegatten zu wählen." (1967)	"Gleicher Lohn für gleichwertige Arbeit. Diese gerechte Forderung bleibt aktuell. [...] Nun ist es aber so, dass gerade in manchen Gebieten eine gleichwertige aber unterschiedlich bezahlte Arbeit besteht. [...] Und im Unterrichtswesen, in den Banken, in den öffentlichen und privaten Verwaltungen, da finden wir in derselben Schule, am selben Pult, gleich qualifizierte Männer und Frauen, die die gleiche Arbeit ausführen, aber die Frau wird weniger bezahlt. [...] In den Banken betragen die Besoldungen bei gleichwertiger Arbeit ein Viertel weniger; zahlreiche Frauen sind als Prokuristen tätig, ohne den Titel zu führen, und ohne die entsprechende Entlohnung, ohne Pension. [...] Die bundesrätliche Botschaft spricht von einer Evolution, und nicht ohne Grund hat der Bund Schweizer Frauenvereine an die nationalrätliche und an die ständerätliche Kommission eine Eingabe gerichtet, die die Ratifikation des Übereinkommens Nr. 100 empfiehlt." (1960)	"Neuenburg verwirklicht ein Frauenpostulat. Prämiengleichheit in der Krankenversicherung in Sicht. [...] Nach dem Entwurf für das revidierte eidgenössische Krankenversicherungsgesetz hätten die Krankenkassen ermächtigt werden sollen, von den Frauen bis zu 25% höhere Prämien zu verlangen als von den Männern. [...] Das neuenburgische Einführungsgesetz bestimmt, dass derjenige Teil der Prämien, welcher vom Versicherten zu entrichten ist, für beide Geschlechter der gleiche sein wird. Der Kanton übernimmt die Hälfte der Gesamtsumme aller Prämien plus einen Zuschlag zum Ausgleich zwischen Männer- und Frauenprämien. [...] Die Aussichten für eine Annahme sind gut, haben doch die Frauen im Kanton Neuenburg seit 1959 das Stimmrecht." (1964)
Dictionary	Frau, Frauen, weiblich, weib, Damen, Dame, Fräulein, Tochter, Mädchen	Frauenstimmrecht, Frauenwahlrecht, Gleichstellung, Frauenbeteiligung, Wahlbeteiligung, Bürgerrecht, Frauenrecht, Frauenstimmund, Frauenstimm, Stimmberechtigten, Schweizerfrauen, Stimmregister, Stimmrecht, Rechtsfähigkeit, Wahlrecht, Kantonsebene, Stimmberechtigte, Ausübung, Bundesebene, Rechtsgleichheit, Demokratie, Verfassung, Schweizerfrau, Mitbestimmung	angeboren, Bürgerrecht, verheiratet, Einbürgerung, ausländische, Ehefrau, Schweizerfrau, Schweizerfrauen, Frau, Frauen, Rechtsgleichheit, Rechtsungleichheit, Heirat, Eheschliessung, Nationalität, Saisonstatus, Schweizerinnen	Lohngleichheit, gleichwertiger, Arbeitsplatz, Entlöhnung, Lohn, Frau, Frauen, Ehefrau, weiblich, Mädchen, Schülerinnen, Diskriminierung, Lohndiskriminierung, Gleichstellung, diskriminierte, Schulbildung, Mädchenbildung, Bildungsmöglichkeit, Hauswirtschaftsunterricht, Haushaltsunterricht, Mathematikunterricht, Naturwissenschaften, Schulprogramm	Mutterschaft, Mutterschaftversicherung, Prämiengleichheit, Prämiengestaltung, gebären, Krankenversicherungsobligatorium, Obligatorium, obglitatorsiche, Frau, Frauen, Ehefrau, weiblich, Unfallversicherung, Hausfrau, Haushaltsunfälle, Haushaltsunfall, Hinterlassenenrente, kinderlos, Witwen, Witwenrente
Embed2discover parameters:					
Step 1: Threshold for classifier confidence	0.5	0.9	0.9	0.9	0.9
Step 1: Remove frequent words	0.1	0.1	0.1	0.1	0.1
Step 2: Min/Max cluster size	3/1000	3/1000	3/1000	3/1000	3/1000
Number of coding rounds	35	40	7	42	56
Number of 'strongly relevant sentences'	130	120	60	110	150
PR accuracy (20-fold)	0.73	0.66	0.75	0.70	0.71
Total bills identified	12,521	157	252	2032	395

Figure 12 offers descriptive insights into the evolution of women-centric bills over time in the Swiss parliament. Figure 12 **A** illustrates the absolute number of women-centric bills submitted to the Swiss parliament over various legislative periods, starting from 1905 to 2020. The bill counts are aggregated over these periods, and questions are not excluded from the count. We observe a significant increase in the number of women-centric bills submitted over time, especially noticeable from the 1980s onwards. The period after 1991 shows a marked increase in the submission of such bills, with the highest number of submissions occurring in the most recent legislative periods. We further highlight three main occurrences, the failed attempt to introduce womens' rights in 1959, the introduction of women's suffrage in 1971 at the Swiss national level, and the first large women's strike in 1991, which was the largest public mobilization of women in Switzerland since 1918. We further observe that this occurrence coincides with a strong increase in submitted women-centric bills in absolute terms. Figure 18 in the Appendix introduces a timeline of the most relevant changes and occurrences in Switzerland with regard to the introduction of women's suffrage.

Figure 12 **B** illustrates the percentage of women-centric bills over the same time periods as Plot A, highlighting significant change points. In the 1940s, women-centric bills mainly dealt with family rights. During the 1960s and 1970s, the focus of these bills shifted to issues concerning marriage, maternity, and insurance discrimination. After 1991, there was a notable increase in bills addressing societal challenges that disproportionately affect women. Several significant change points are marked throughout this period, indicating instances where the percentage of women-centric bills significantly increased.

Figure 12 **C** represents the number of bills submitted over time that relate to **specific political demands** of Swiss women. Early legislative efforts focused on securing the right to vote and ensuring equal representation in all governing bodies. In 1984, the demand for the preservation of citizenship upon marriage was granted, with subsequent bills aiming at the re-attainment of lost citizenships. More recently, there has been an increased focus on the demand for equal pay and the eradication of wage discrimination. Additionally, there is a notable number of bills addressing education opportunities for women, insurance discrimination, and maternity leave.

5.2 Effects of women's enfranchisement on political engagement of Swiss members of parliament

Over time, there has been a notable increase in both the percentage and the absolute number of women-centric bills and those addressing the political demands of Swiss women.

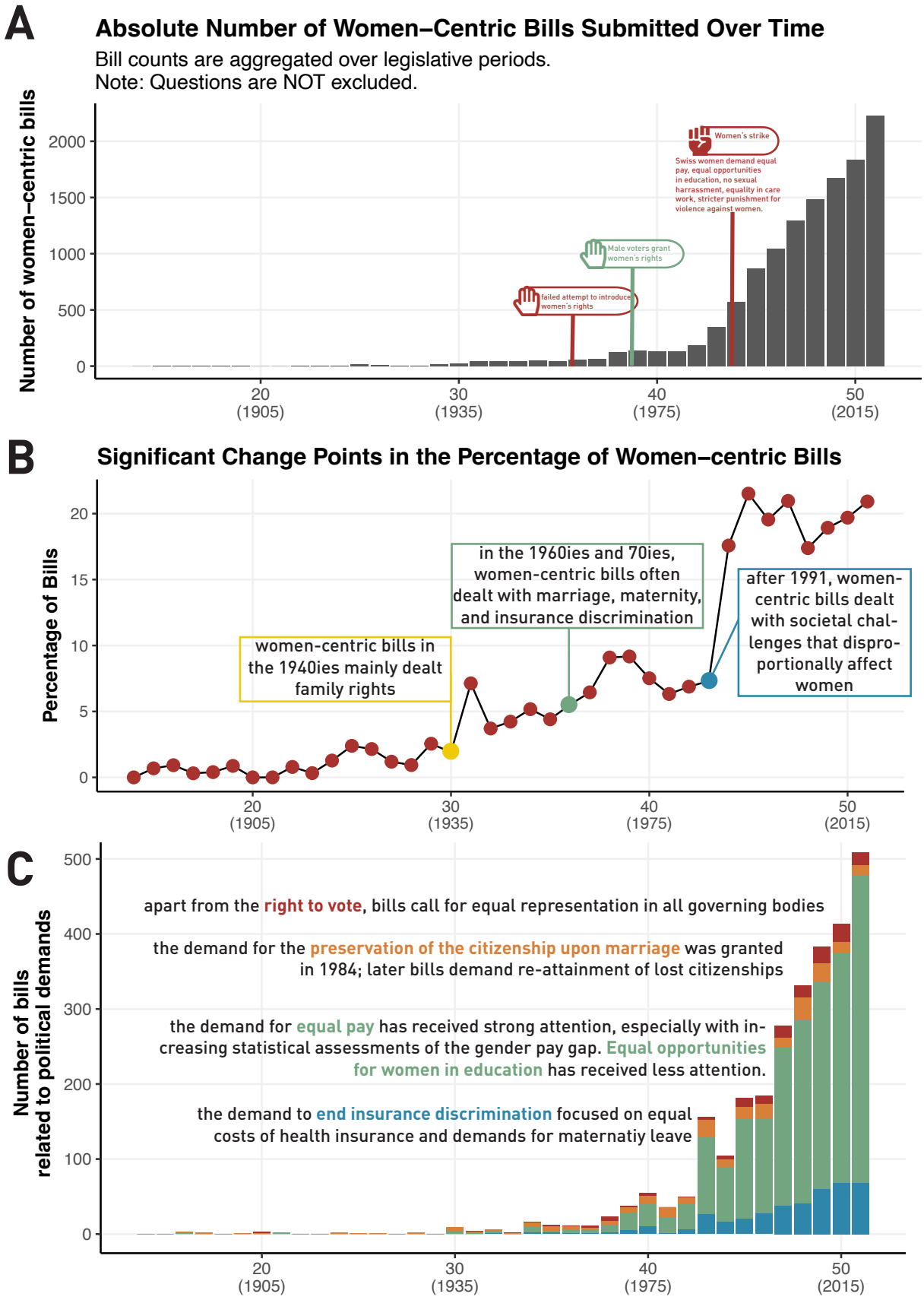
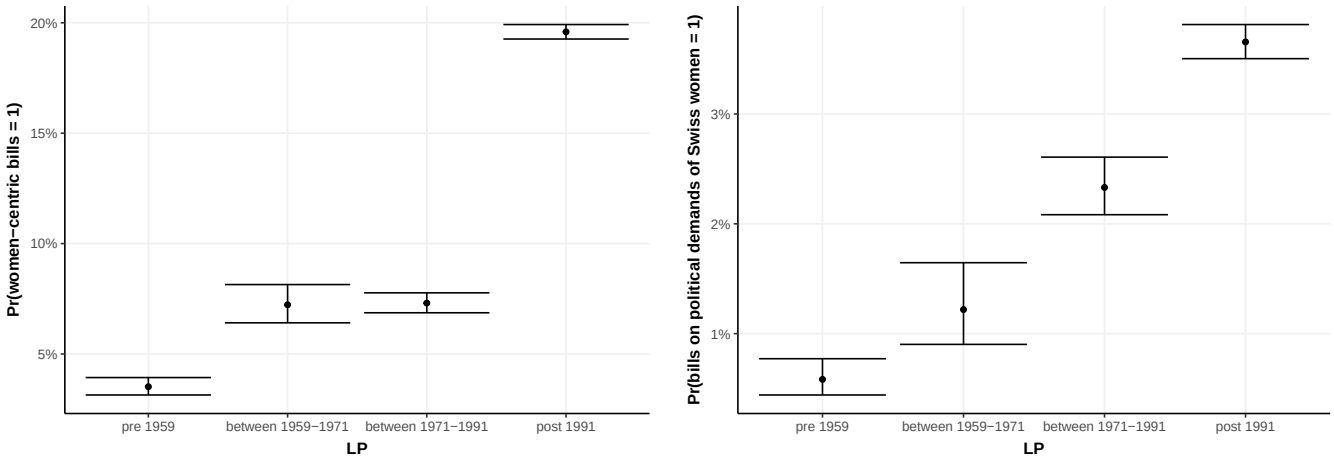


Figure 12: Descriptive insights of women-centric bills over time. **A** illustrates the increase of women-centric bills over time in absolute numbers. **B** indicates three significant change points in the submitted bills. **C** shows the increase of bills for each of the specific political demands.



(a) Logistic regression on **women-centric** bills

(b) Logistic regression on **women's demand** bills

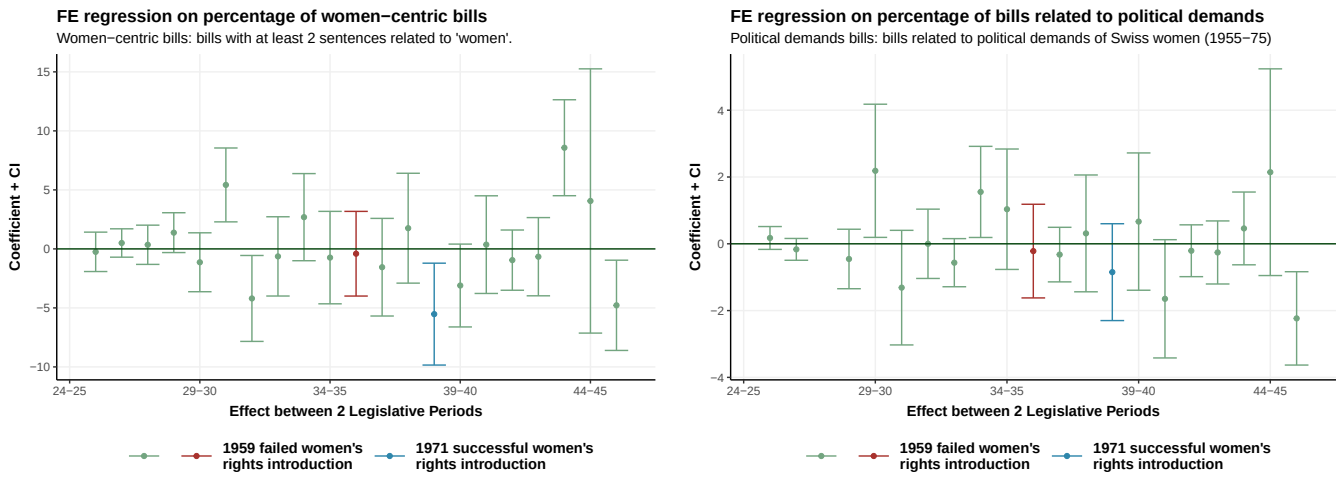
Figure 13: Logistic regression on women-centric and women's demand bills

Figures 13(a) and 13(b) illustrate the results of a logistic regression on whether a submitted bill is a women-centric bill, or a women's demands bill, respectively. In Figure 13(a), we observe a notable increase in the post-1991 period, reaching approximately 20%. In contrast, the earlier periods, specifically pre-1959 and between 1959-1971, exhibit a lower probability of a bill being women-centric. This indicates a significant shift in legislative focus towards women's issues in the later periods.

Figure **??** depicts a similar trend to Figure 13(a). However, the increasing trend is more gradual over the selected time periods. As the parliament tackles a broad variety of issues, the fraction of bills addressing specific women's demands is small, as is reflected in the small marginal effects. Nonetheless, over time, the probability quadruples between the pre-1959 period and the post-1991 period.

Figures 14(a) and 14(b) present marginal effects from two-way fixed effects regression models on the percentage of women-centric bills, and women's demands bills respectively, showing the effects between two legislative periods and fixing on within-MP effects. Two key historical events are highlighted: the failed women's rights introduction in 1959 and the successful introduction in 1971. The coefficients and confidence intervals are depicted for each comparison between legislative periods. We see that MPs who experienced the 1971 enfranchisement, did not change their behavior. In fact, MPs tend not to change their engagement, no matter the times.

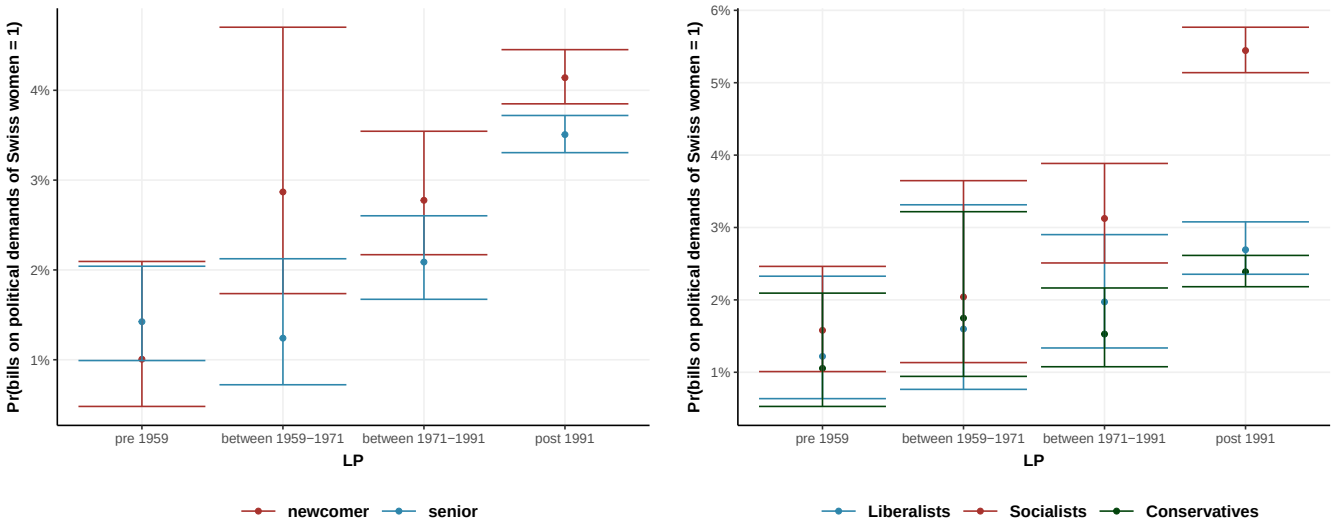
Given these insights, we thus turn to the question of ***who*** submits those women-demands bills. Figure 15(a) shows that in particular post 1991, it is more likely that a women-demands bill is submitted by a



(a) FE Regression on Percentage of Women-Centric Bills (b) FE Regression on Percentage of Bills Related to Political Demands

Figure 14: FE Regression on Percentage of Women-Centric Bills and Bills Related to Political Demands

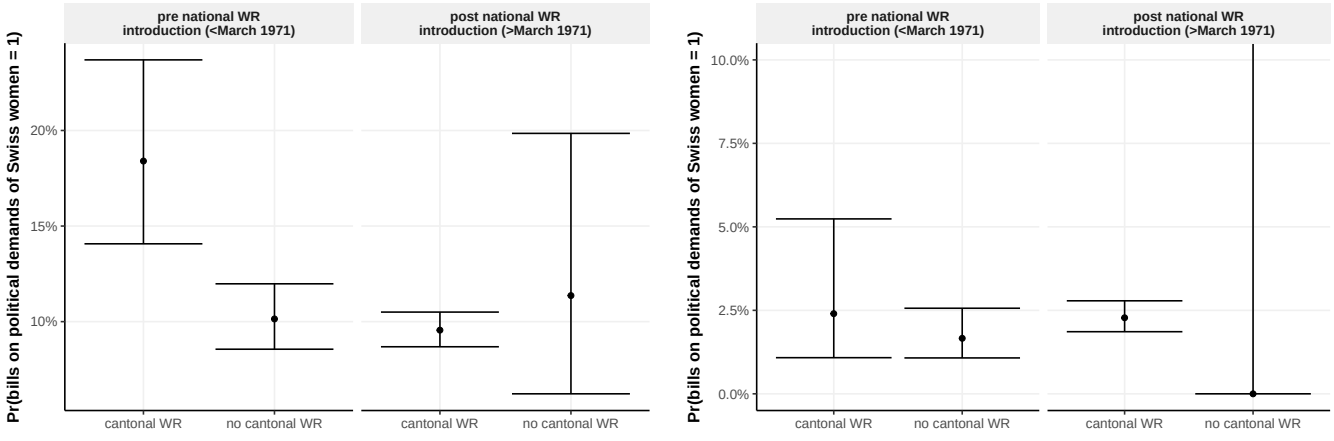
newcomer MP. Figure 15(b) shows that in particular post 1991, it is more likely that a women-demands bill is submitted by a *socialist* MP. Figures 16 show that between 1959 and 1971, it is more likely that a women-demands bill is submitted by an MP that represents a Swiss *canton that already granted women their right to vote* on the cantonal level. Finally, Figure 17 shows that in particular right after the national enfranchisement in 1971, it is more likely that a women-demands bill is submitted by a *female* MP.



(a) Newcomers versus senior MPs

(b) Party family affiliation of MPs

Figure 15: Marginal effects of newcomers (a) and party family affiliation (b) on bills being women-centric or tackling a women's demand. Effects are split over 4 time periods.



(a) Women-centric bills

(b) Women's demands bills

Figure 16: Marginal effects of cantonal differences on bills being women-centric (a) or tackling a women's demand (b).

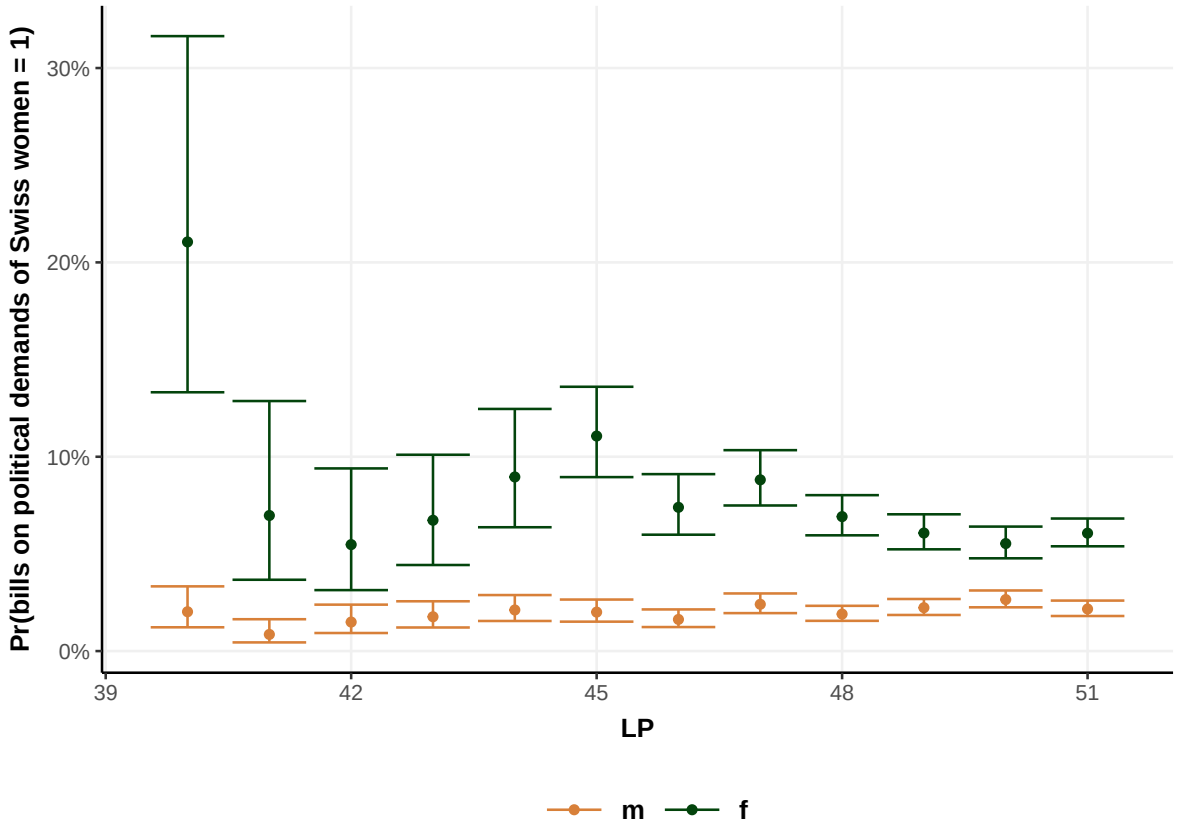


Figure 17: Marginal effects of gender differences on bills tackling a women's demands.

6 Conclusion

This paper introduces `embed2discover`, a semi-automated tool designed for dictionary-based content analysis, combining advanced natural language processing techniques with human annotations. This paper demonstrates how `embed2discover` can efficiently expand and refine dictionaries to accommodate linguistic and contextual changes over time, ensuring comprehensive and accurate text classification for large corpora. Overall, `embed2discover` presents a significant advancement in the field of automated content analysis, bridging the gap between machine efficiency and human interpretative capabilities. Additionally, this paper provides two showcases that highlight `embed2discover`'s versatility: one focusing on the evolution of biodiversity-related bills and the other on legislative elites' responses to women's enfranchisement in Switzerland. With the help of `embed2discover`, we were able to code broad topics as well as very narrow women's demands. These applications underscore the potential of `embed2discover` to enhance traditional content analysis and classification methods, providing more accurate and nuanced insights into large-scale text corpora.

7 Appendix

7.1 Timeline of women's suffrage in Switzerland: major milestones

Figure 18 provides an overview of the long road to women's suffrage in Switzerland.

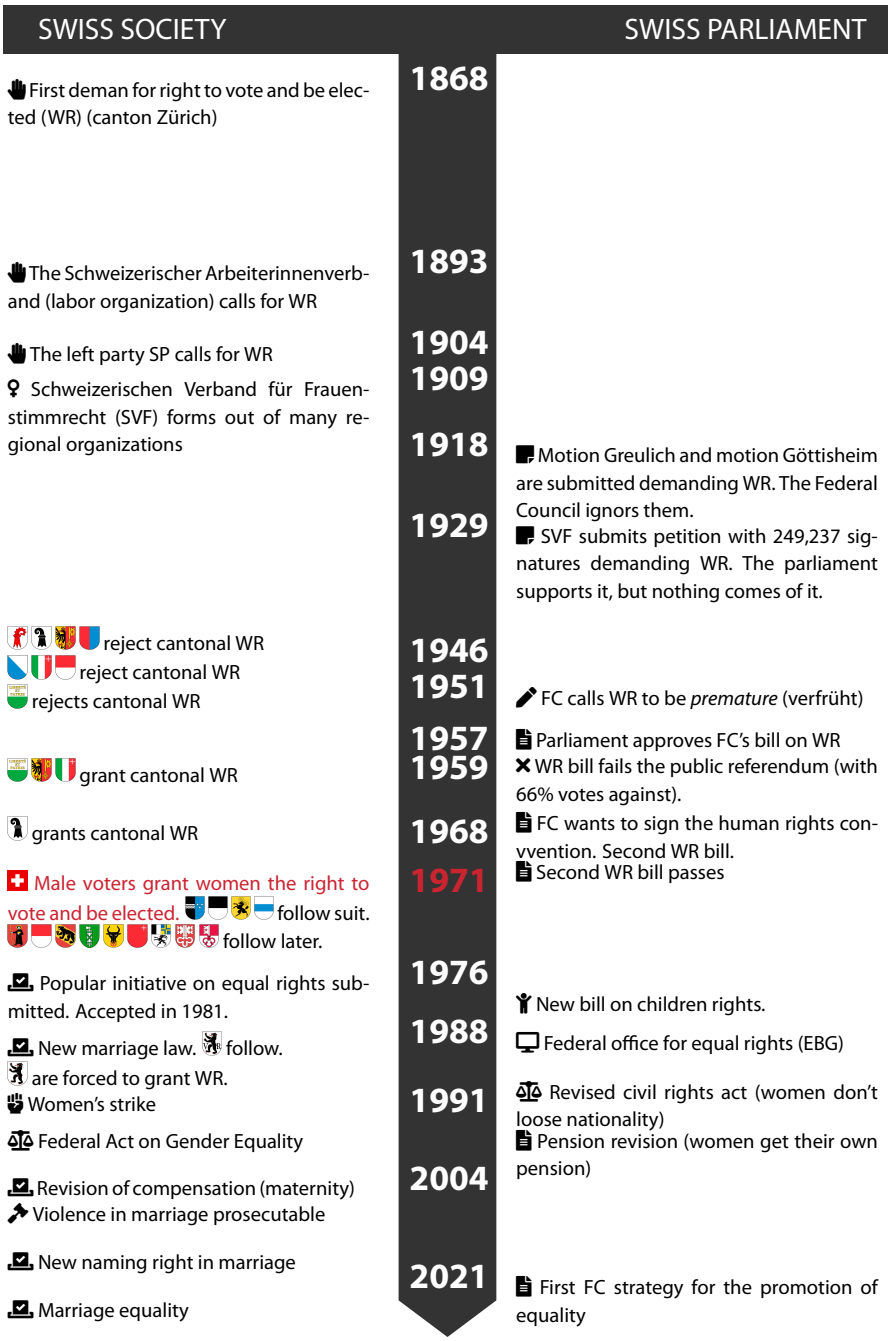


Figure 18: Timeline of major women's suffrage milestones in Switzerland. The arms represent the respective canton of Switzerland.

References

Acemoglu, Daron and James A Robinson. 2000. "Political losers as a barrier to economic development." *American Economic Review* 90(2):126–130.

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. pp. 2623–2631.

Andersen, Peggy M, Philip J Hayes, Steven P Weinstein, Alison K Huettner, Linda M Schmandt and Irene Nirenburg. 1992. Automatic extraction of facts from press releases to generate news stories. In *Third conference on applied natural language processing*. pp. 170–177.

Baden, Christian, Christian Pipal, Martijn Schoonvelde and Mariken AC G van der Velden. 2022. "Three gaps in computational text analysis methods for social sciences: A research agenda." *Communication Methods and Measures* 16(1):1–18.

Barberá, Pablo, Amber E Boydstun, Suzanna Linn, Ryan McMahon and Jonathan Nagler. 2021. "Automated text classification of news articles: A practical guide." *Political Analysis* 29(1):19–42.

Berlinski, Samuel, Torun Dewan et al. 2011. "The political consequences of franchise extension: Evidence from the second reform act." *Quarterly Journal of Political Science* 6(34):329–376.

Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov. 2016. "Enriching Word Vectors with Subword Information." *arXiv preprint arXiv:1607.04606* .

Bommasani, Rishi, Kelly Davis and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 4758–4781.

Budge, Ian. 2001. *Mapping policy preferences: estimates for parties, electors, and governments, 1945-1998*. Vol. 1 Oxford: Oxford University Press.

Carley, Kathleen. 1990. "Content analysis." *The encyclopedia of language and linguistics* 2:725–730.

Carley, Kathleen. 1994. "Extracting culture through textual analysis." *Poetics* 22(4):291–312.

Chatsiou, Kakia and Slava Jankin Mikhaylov. 2020. "Deep learning for political science." *The SAGE handbook of research methods in political science and international relations* pp. 1053–1078.

Corder, J Kevin and Christina Wolbrecht. 2006. "Political context and the turnout of new women voters after suffrage." *Journal of Politics* 68(1):34–49.

Cowie, Jim and Wendy Lehnert. 1996. "Information extraction." *Communications of the ACM* 39(1):80–91.

Dai, Yaoyao and Alexander Kustov. 2022. "When do politicians use populist rhetoric? Populism as a campaign gamble." *Political Communication* 39(3):383–404.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* .

Fisher, Dana R, Joseph Waggle and Philip Leifeld. 2013. "Where does political polarization come from? Locating polarization within the US climate change debate." *American Behavioral Scientist* 57(1):70–92.

Franzosi, Roberto, Gianluca De Fazio and Stefania Vicari. 2012. "Ways of measuring agency: an application of quantitative narrative analysis to lynchings in Georgia (1875–1930)." *Sociological Methodology* 42(1):1–42.

Greene, Kevin T, Baekkwan Park and Michael Colaresi. 2019. "Machine learning human rights and wrongs: How the successes and failures of supervised learning algorithms can inform the debate about information effects." *Political Analysis* 27(2):223–230.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(03):267–297.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24:395–419.

Hanna, Alexander. 2013. "Computer-aided content analysis of digitally enabled movements." *Mobilization: An International Quarterly* 18(4):367–388.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, Adriane Boyd et al. 2020. "spaCy: Industrial-strength natural language processing in python.".

Jagers, Jan and Stefaan Walgrave. 2007. "Populism as political communication style: An empirical study of political parties' discourse in Belgium." *European journal of political research* 46(3):319–345.

Jaligot, Rémi, Jérôme Chenal, Marti Bosch and Stéphanie Hasler. 2019. "Historical dynamics of ecosystem services and land management policies in Switzerland." *Ecological Indicators* 101:81–90.

Johnson, Jeff, Matthijs Douze and Hervé Jégou. 2019. "Billion-scale similarity search with GPUs." *IEEE Transactions on Big Data* 7(3):535–547.

Jordan, Soren, Hannah L Paul and Andrew Q Philips. 2023. "How to cautiously uncover the "Black Box" of machine learning models for legislative scholars." *Legislative Studies Quarterly* 48(1):165–202.

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American political science Review* 107(2):326–343.

Klingemann, Hans-Dieter. 2006. *Mapping policy preferences II: estimates for parties, electors, and governments in Eastern Europe, European Union, and OECD 1990-2003*. Vol. 2 Oxford: Oxford University Press.

Laurer, Moritz, Wouter Van Atteveldt, Andreu Casas and Kasper Welbers. 2024. "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI." *Political Analysis* 32(1):84–100.

Lee, Sungjick and Han-joon Kim. 2008. News keyword extraction for topic tracking. In *2008 fourth international conference on networked computing and advanced information management*. Vol. 2 IEEE pp. 554–559.

McConnaughy, Corrine M. 2013. *The woman suffrage movement in America: A reassessment*. Cambridge: Cambridge University Press.

Mikhaylov, Slava, Michael Laver and Kenneth R Benoit. 2012. "Coder reliability and misclassification in the human coding of party manifestos." *Political Analysis* 20(1):78–91.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *Proceedings of Workshop at the International Conference on Representation Learning*. pp. 1–12.

Morgan-Collins, Mona. 2021. "The electoral impact of newly enfranchised groups: The case of women's suffrage in the United States." *The Journal of Politics* 83(1):150–165.

Muddiman, Ashley, Shannon C McGregor and Natalie Jomini Stroud. 2019. "(Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries." *Political Communication* 36(2):214–226.

Nelson, Laura K, Derek Burk, Marcel Knudsen and Leslie McCall. 2021. "The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods." *Sociological Methods & Research* 50(1):202–237.

Nussio, Enzo and Govinda Clayton. 2024. "Introducing the lynching in Latin America (LYLA) dataset." *Journal of Peace Research* pp. 1–18.

Reimers, Nils and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* .

Sahlgren, Magnus. 2008. "The distributional hypothesis." *Italian Journal of linguistics* 20:33–53.

Salamanca, Luis, Laurence Brandenberger, Lilian Gasser, Sophia Schlosser, Marta Balode, Vincent Jung, Fernando Perez-Cruz and Frank Schweitzer. 2024. "Processing Large-Scale Archival Records: The Case of the Swiss Parliamentary Records." *Swiss Political Science Review. Online First.* pp. 1–34.

Schlosser, Sophia, Laurence Brandenberger, Julian Minder, Giuseppe Russo, Luis Salamanca and Frank Schweitzer. 2023. "From Expertise to Versatility: The Evolution of Issue Engagement in the Swiss Parliament Over 130 Years." *Paper presentation at the European Political Science Association 13th Annual Conference, EPSA, Glasgow, UK, June 22-24, 2023.* pp. 1–12.

Skorge, Øyvind Søraas. 2023. "Mobilizing the underrepresented: Electoral systems and gender inequality in political participation." *American Journal of Political Science* 67(3):538–552.

Song, Hyunjin, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga and Hajo G Boomgaarden. 2020. "In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis." *Political Communication* 37(4):550–572.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The ties that double bind: social roles and women's underrepresentation in politics." *American Political Science Review* 112(3):525–541.

Vamvas, Jannis, Johannes Graën and Rico Sennrich. 2023. "SwissBERT: The Multilingual Language Model for Switzerland." *ArXiv Preprint:2303.13310* pp. 1–15.

Van Atteveldt, Wouter, Mariken ACG Van der Velden and Mark Boukes. 2021. "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms." *Communication Methods and Measures* 15(2):121–140.

Vourvachis, Petros and Thérèse Woodward. 2015. "Content analysis in social and environmental reporting research: trends and challenges." *Journal of Applied Accounting Research* 16(2):166–195.

Wilkerson, John and Andreu Casas. 2017. "Large-scale computerized text analysis in political science: Opportunities and challenges." *Annual Review of Political Science* 20(1):529–544.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz et al. 2019. "Huggingface's transformers: State-of-the-art natural language processing." *arXiv preprint arXiv:1910.03771* .

Zhang, Yifan, Peilin Zhao, Jiezhang Cao, Wenye Ma, Junzhou Huang, Qingyao Wu and Mingkui Tan. 2018. Online adaptive asymmetric active learning for budgeted imbalanced data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. pp. 2768–2777.