University of Cologne


# Department of Information Systems and Information Management

Prof. Dr. Detlef Schoder

***Master Thesis:***
*Topic Modeling in Personalized Radio*

supervised by
Dr. Stefan Hirschmeier

by Tom Celig

## Table of Contents

## List of Figures

## List of Tables

## Abbreviations

| | |
|---|---|
| API | Application Programming Interface |
| BERT | Bidirectional Encoder Representation of Transformers |
| CBOW | Continuous Bag of Words |
| CPU | Central Processing Unit |
| GPU | Graphics Processing Unit |
| LDA | Latent Dirichlet Allocation |
| MMR | Maximal Marginal Relevance |
| NLP | Natural Language Processing |
| NSP | Next Sentence Prediction |
| TF-IDF | Term Frequency – Inverse Document Frequency |
| UI | User Interface |
| UMAP | Uniform Manifold Approximation and Projection |

# 1 Introduction

A growing number of companies are in need of metadata (Hirschmeier et al., 2019). The application of machine learning algorithms and progress in Natural Language Processing (NLP) through word embeddings and keyword identification can be named driving forces behind this development (Hirschmeier et al., 2019). A well-developed categorization system is of the highest importance because it forms the basis for many applications, such as content-based recommender systems. They enable comprehensive data structuring, which allows customers and companies alike to navigate big datasets. Companies utilize taxonomies internally, which in many cases grew historically based on experts' opinions. They are used in internal categorization tasks. With increasing amounts of data and changes in culture, companies continue to open up and enable users to navigate their data libraries. It is uncertain whether those internal taxonomies are fit for the changing culture and external use by customers. Additionally, the growing amounts of data make manual processing increasingly time-consuming, and the need for automation arises.

Contrary to information retrieval, when users are aware of what they are looking for in some cases, such as media- streaming and library contexts, they need to get an abstract idea of the themes to understand the collection first. Based on this understanding of the corpus, the documents of interest can be explored (Hu et al., 2014). Furthermore, a quality set of potential taxonomy keywords and categories is necessary to enable accurate and meaningful data classification (Zhang et al., 2018).

Topics have multiple characteristics that need to be considered when developing a system. First of all, they can be divided into multiple subcategories. Second, topics are continuous because they can be described by a collection of weighted words that represent each topic; those words can occur in different topics (Angelov, 2020). The derivation of said categories can be described as topic modeling: the aim to find short descriptions of documents that enable the processing of datasets while preserving essential statistical relationships (Blei et al., 2003) by abstracting and summarizing the information on a high level.

There are many ways in which category systems are developed that have certain advantages and disadvantages. The first option is the historical growth of the system and the gradual development of a taxonomy. In the context of a company that

manages a media library, this may be the addition of new terms based on content that is added over time. The major disadvantage of this approach is that the system might turn out structurally inconsistent and may lack logical consistency.

The second option is the complete definition of a categorical system from scratch based on an expert's opinion. Shortfalls, in this case, might include subjective categorization and highly skewed distributions in the metadata (Hoxha et al., 2016). Both approaches share the fact that the categories are not derived automatically but manually by an expert.

In commercial contexts, document corpora rarely stay finite, and new documents are added daily. Therefore, gradual shifts of interest might occur, and the domain-specific taxonomies become outdated. Thus, regular reiterations of topic modeling need to occur, resulting in a resource-intensive task when conducted manually by experts. This fact results in a growing interest in fully and semi-automatic topic modeling approaches (Carrion et al., 2019). Nonetheless, it needs to be evaluated whether the algorithmic approaches can satisfy all requirements and produce coherent, holistic topics as an expert might derive manually. To explore the field of topic modeling further, ways of deriving categorization systems focusing on data-driven approaches need to be developed and compared to the given methods to evaluate which works best in the given use case. Ways to achieve this task, like topic modeling approaches (LDA), do already exist but have not been thoroughly evaluated in a commercial application. Furthermore, the experience on the customer side should be improved to get an optimal mapping of the user's interest profile.

This leaves us with the following research question: *How do users perceive different topic modelings to specify their interests?*

To answer this question, we will compare three different taxonomy systems, a manual, semi-automatic, and fully automatic topic modeling approach, which will be implemented in a prototype and presented to users in the course of a user study.

Multiple challenges need to be faced. First of all, due to the unsupervised nature of automated topic models, objective evaluation poses a challenge. The domain-specific nature of taxonomies and the lack of specific evaluation methods for the coherence and accuracy of unsupervised systems require special consideration. Second, in commercial contexts, the system is required to be comprehensive on the customer's side. The structuring and the degree of coverage of overall topics in the

document corpus need to be determined. The system could cover all potential interests for a customer in the media library context to like or focus on a few overarching topics.

Furthermore, it needs to be evaluated if the categories need to represent the data adequately. If there is a high diversity in subcategories of a particular topic in the document corpus, the taxonomy could reflect this to increase classification specificity potentially. An equal weighting of subcategories, on the other hand, may provide a more holistic, unbiased picture of the potential topics on the customer side. Finally, a significant challenge is the granularity of the categories. If the division is too fine, there might be too much choice for the end-user. Additionally, classifications might turn out inaccurate due to the high specificity of the categories. A too coarse division, though, does not allow the user to specify his interests accurately enough.

This paper starts by laying the foundations and explaining the theoretical background for category construction in general and various Topic Modeling approaches. Next, the methodology applied in the research paper is explained, including metrics and frameworks used. In the "Implementation" chapter, the concrete implementation of the automated topic models is explained, such as preprocessing, hyperparameter optimization, and resulting findings. In the following chapter, the results of the user study are presented and subsequently discussed. Finally, the presentation of limitations that were faced, as well as the conclusion, takes place.

# 2 Theoretical Background

In the following chapter, the theoretical background for the various topic modeling techniques will be explained. Carrion divides the techniques into manual, semi-automatic, which require some human input, and fully automatic, which require no human input (Carrion et al., 2019). We will only reference, for the sake of simplicity, manual and automatic methods. The distinction between semi-automatic and fully automatic is a matter of interpretation because there is always some form of human intervention involved, for example, in data selection.

## 2.1 Manual Topic Modeling

One of the main approaches to derive category systems and taxonomies from unstructured data is the application of manual techniques. Those can include, for example, domain experts structuring the data based on their knowledge and experience in the given field. Depending on the requirements, they may develop the categories from scratch or recycle and adapt existing approaches (Carrion et al., 2019). A significant benefit of the manual method is that resulting taxonomies are well structured and easily comprehensible by humans compared to automatic techniques that often produce noisy and challenging to interpret results (Kotlerman et al., 2011). The downside is the high time consumption and manual effort required to process a large document corpus. Furthermore, a high level of expertise is necessary.

In general, we can divide category construction into two primary approaches. The first one is a *top-down* approach and the second one being *bottom-up*. The *top-down* approach works with apriori assumptions and is often guided by a framework or set for predefined terms in the form of a taxonomy. Themes and categories are known beforehand by the researcher and mapped onto the data. A predefined coding schema can be derived from literature or experts (Urquhart, 2012). A *bottom-up* approach, on the other hand, does not work with prior assumptions or frameworks. No preconceptions should be imposed on the data, and the analysis of topics should start on a word and sentence level and then aggregated (Urquhart, 2012).

Both methods have their benefits and disadvantages. The top-down method may provide a more structured system and lead to a result grounded in literature and based on a given schema. On the other hand, the bottom-up approach will identify data-specific concepts and minimize the chance of missing an important category

(Urquhart, 2012). During practical application, researchers also combine both concepts to utilize the benefits of both methods. The choice of method is highly circumstantial and dependent on multiple variables. The availability of resources and size of the dataset needs to be considered when considering a manual bottom-up analysis. Working on a sentence level requires more effort than a top-down analysis.

Furthermore, the desired result needs to be considered. A bottom-up approach is suitable if all thematic concepts down to a granular level need to be found. In contrast, a top-down approach is preferable when structural requirements or predefined schemata are of importance.

## 2.2 Data-Driven Topic Modeling

Methods like clustering and dimensionality reduction are called unsupervised methods as they require no human intervention and are not dependent on human knowledge. Furthermore, they enable us to work on large datasets, thus saving considerable amounts of time. Nonetheless, the results produced by unsupervised methods need to be post-processed and cleaned to conceptualize the results produced by the algorithm. Unsupervised Topic Modeling approaches belong to the bottom-up methods as they attempt to find features and categories inductively (Debortoli et al., 2016). In the following, three unsupervised approaches will be explained, which were utilized as a part of the study.

### 2.2.1 LDA

A prominent unsupervised approach to topic modeling and text document labeling is Latent Dirichlet Allocation (LDA). Usually, it is used in dimensionality reduction during feature analysis, but it can also reduce texts down to keywords and topics. The model assumes that each document is a bag of words with a set of topics. By analyzing the distribution of words throughout all documents, the model determines which words form a topic. If multiple documents contain the same unique subset of words, it increases the probability that this combination of words belongs to the same topic. After getting a predefined number k of topics, the model then calculates the probability distribution of each document belonging to a specific topic based on the words contained within the document. Finally, it returns the probability of topics

for each document and the distribution of words for each topic (Blei et al., 2003). A significant disadvantage of LDA is that the model assumes the number of topics k to be known. Especially in large and unknown datasets, this is rarely the case. Furthermore, the model lacks semantic understanding as it is based on the bag of words approach. For example, *Canada* and *Canadian* would be treated as different words (Angelov, 2020).

### 2.2.2 Top2Vec

In 2013 Mikolov et al. developed an unsupervised approach to convert words into a distributed spatial vector representation (word2vec). It accepts a corpus (document, sentence) and loops over each term, and tries to predict adjacent terms or vice versa, meaning its context. The output is a vector for each term containing information regarding its context. The significant advantage to prior keyword identification approaches is that semantic relationships of a word are still contained within the spatial shape and geometrical distance (Mikolov et al., 2013). There are two different versions of the two-layered model network that can be utilized, CBOW (Continuous Bag Of Words) and Skip-Gram. The first version of the algorithm employs a moving window and tries to predict each term in the window based on the other terms, while the inverse is true for the Skip-Gram version: the adjacent terms are predicted based on a single word (Mikolov et al., 2013).

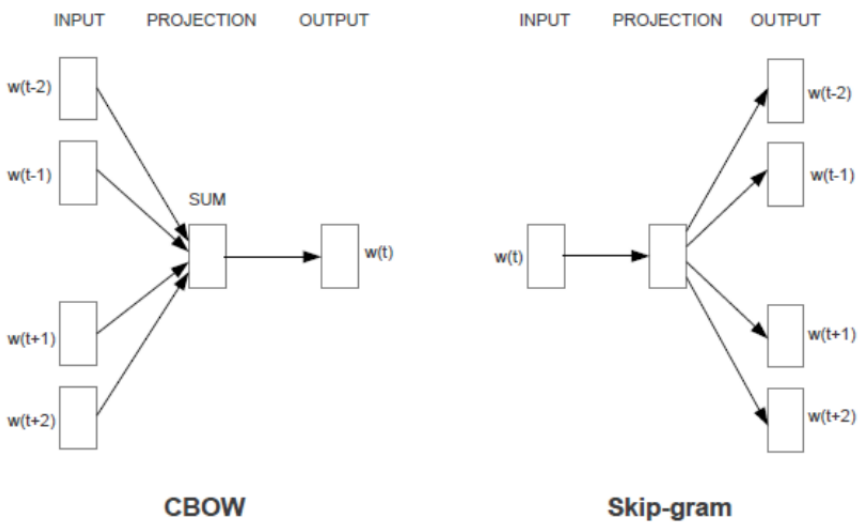


Figure 1: CBOW and Skip-gram

In 2014, they expanded the approach to whole paragraphs and text documents, enabling the representation of paragraphs as single word vectors capturing the content. Paragraph to Vec (doc2vec) works by "concatenating the paragraph vector with several word vectors from a paragraph and predicting the following word in the given context" (Le & Mikolov, 2014). The additional vector is also called the paragraph ID, which is document unique (Fig. 2). Therefore, by training the word vectors in the same manner as in word2vec, the paragraph ID vector will contain an abstract, unique representation of the given document or paragraph as it was trained alongside all words in the document. This method enables us to capture the overall logical structure of the documents.

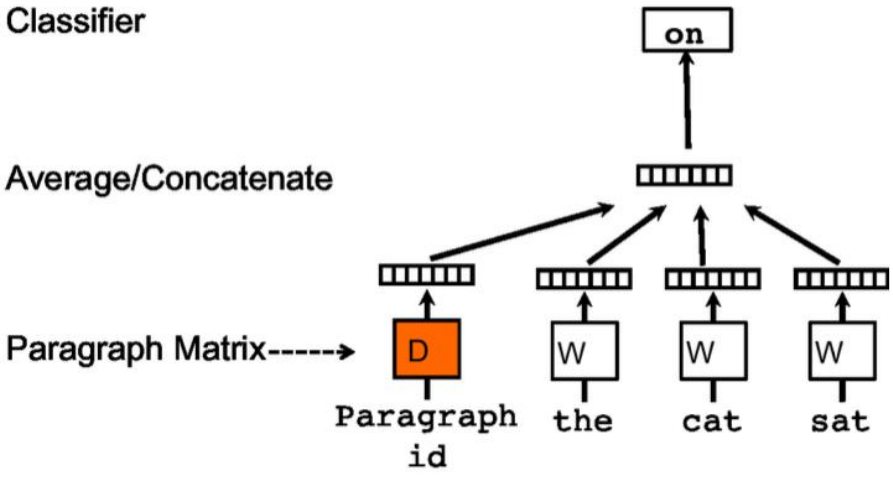


Figure 2: Doc2Vec visualized (Le & Mikolov, 2014)

Like word2vec, doc2vec also exists in the CBOW and Skip-Gram variants. Based on Mikolovs work, Angelov proposes a model called top2vec, which produces embedded topic vectors. They are produced by calculating dense areas in the document corpus's semantic space. This is achieved by deriving document vectors for each document via doc2vec first. After the vectors are derived, UMAP (Uniform Manifold Approximation and Projection) is applied to reduce their dimension. This step aids with the reduction of computing resources and can increase performance when working with density-based clustering algorithms. UMAP uses a graph layout system to project data into a lower-dimensional system. A low-dimensional graph is optimized until it resembles the high-dimensional data as closely as possible. This leads to

achieving high-performance scores compared to prior dimensionality reduction methods by keeping a large portion of the higher-dimensional structures in the lower-dimensional space (Angelov, 2020).

Finally, the reduced document vectors are clustered via HDBSCAN. It is a fitting clustering algorithm due to the high amount of noise and variability in the density of the clusters. Conventional clustering methods like k-means fail to perform well on data containing clusters with arbitrary shapes, different sizes, densities, or with a lot of noise and outliers. All of these properties apply to our semantic space. Furthermore, HDBSCAN can determine the number of clusters independently, which is a significant benefit as the number of topics in many cases is not known in advance. The algorithm proceeds to identify dense areas and label documents with either noise labels or labels for the cluster to which it belongs. Dense areas in the semantic space can be interpreted as highly similar documents with a common underlying topic.

To determine topic words, word vectors to each topic vector (cluster centroid) are determined, representing the given document cluster. It is determined by calculating the arithmetic mean of all documents in a given cluster. The authors found that the cluster centroid provides a sufficient representation of the underlying topic in the entire cluster. Every point in the semantic space represents a topic whose semantics are best defined by the nearest word vectors. As a result, the semantically most representative word vectors are those closest to the cluster centroid vector. The distance between them determines the semantic similarity of each word vector to the centroid vector. Thus, the resulting topic words are the word vectors closest to each topic vector (Angelov, 2020).

A benefit of topic vectors and continuous representations compared to prior methods is the dynamic reduction of the number of topics. Smaller topics can be iteratively merged into their closest topic in the semantic space. This procedure can be repeated until the desired number of topics is reached. Topic sizes and vectors then have to be recalculated.

The authors benchmarked top2vec against LDA and PLSA and conducted that the model delivered more informative topic words with more extensive semantic similarities between each other. Mainly because generative models like LDA and PLSA rely on recreating the distribution of the original document with a minimal loss which leads to the inclusion of uninformative words because they make up large

portions of all documents. Furthermore, there is no guarantee that the found topics are representative of the entire document corpus.

### 2.2.3 BERT

Bidirectional encoder models such as BERT (Bidirectional Encoder Representations from Transformers) have been achieving state-of-the-art results in various fields of NLP processing tasks. Mainly, pre-trained models provide the primary benefit of being trained on large datasets, leading to extensive representations of sentences and words (Devlin et al., 2019). These pre-trained models then just have to be fine-tuned for the given task. This addresses significant drawbacks that exist with prior embedding models such as word2vec as the quality and amount of training data is of the highest importance for the model's performance. These training corpora, like the entire Wikipedia, contain a wide range of topics and are of high quality.

BERT is an NLP model based on transformers, which learn contextual and semantic relations between words in a document. Transformers have two different processes in their basic form: an encoder that reads the text input and a decoder that generates a prediction. Only the encoder technique is required for BERT since its purpose is to construct a semantic language model.

For the application, two main steps are to be conducted. First, the model has to be trained on the initial, large corpus (pretraining). Then, it needs to be fine-tuned to the given NLP task to achieve optimal performance. When working with the model, a series of tokens are fed into the BERT encoder, subsequently transformed into vectors, and processed by the neural network. However, before BERT can begin processing, the input must be enhanced with additional metadata. Three types of embeddings are contained in the input:

1. **Token Embeddings**: They contain information on the actual word and stem from the WordPiece token vocabulary.
2. **Position Embeddings**: They mark the position of each word in the sentence. Prior approaches were limited by not being able to capture information on sequence or order.
3. **Segment Embeddings**: For sentence pairs, each sentence is marked with a unique embedding to differentiate between them.

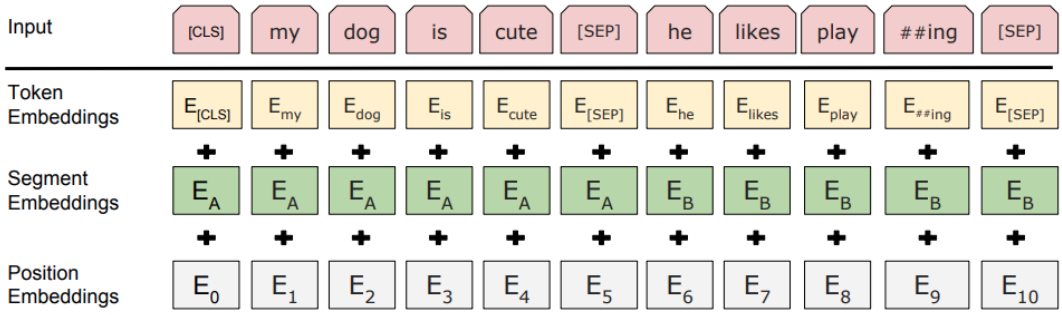


Figure 3: BERT Embeddings (Devlin et al., 2019)

The pretraining is conducted in two steps, Masked Language Modeling and Next Sentence Prediction. Prior language models learned by trying to predict words in sentences sequentially. This imposes a limit in flexibility and semantic understanding. Thus, for the first step, a bidirectional method is introduced to learn the surrounding context for each word. Rather than predicting the next word in a sequence (as in prior approaches), random words in the sequence are masked and predicted based on the surrounding words. This leads to a model with a high, bidirectional understanding of relationships between words. For the second step, the Next Sentence Prediction (NSP), the Segment Embeddings are used. They act as binary classification labels. Based on those, the model is trained to predict whether a sentence in a given sentence pair is preceding or following the other sentence. The resulting model is a model with a high understanding of relationships between words and sentences, thus covering the concepts throughout the entire document.

When applying the BERT model in a specific NLP context, fine-tuning should be conducted to achieve optimal performance. By feeding the data that is worked with as input into the model as loss function is optimized and specifics in the data used in the tasks are reflected in the network. This expands the model from having a general understanding of natural language processing based on the pretraining corpus to a model that is fit to embed and work with the text data used in the project specifically (Devlin et al., 2019).

### 2.2.4 BERTopic

By generalizing the approach described in *top2vec*, it is possible to derive a method that leverages advanced document and word embeddings for topic modeling.

Groostendorst developed an algorithm called BERTopic, implementing this idea. Similar to top2vec, the algorithm functions in three main steps.

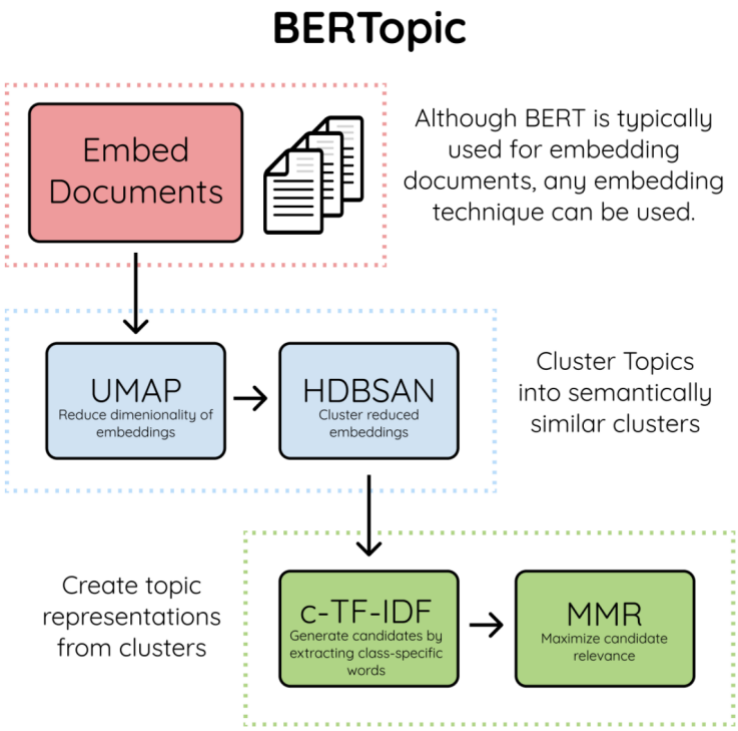


Figure 4: BERTopic visualized (*The Algorithm - BERTopic*, 2021)

Initially, embeddings are extracted from the document corpus with BERT. Even though BERT and variations of BERT are the main algorithms used as they yielded the best results in the approach, the authors designed it so that any other embedding technique can extract the embeddings. BERT variations can include models that feature minor differences in the neural network architecture, such as different pooling layers or base models. Furthermore, the training data is a differentiating factor as it determines which language the model can work on. There are language-specific models and models that are multilingual such as "xlm-r-bert-base-nli-stsb-mean-tokens" which works on over 50 languages. Different approaches thus might be more suited for different data or use-cases.

Next is the clustering portion of the approach. Dimensionality is reduced with UMAP, and the embeddings of semantically similar documents are clustered with HDBSCAN, as was the case in *top2vec*. The dimensionality reduction is not only essential to increase efficiency and reduce computation times but also due to

HDBSCAN being susceptible to the curse of dimensionality, which means that with too many features, the algorithm might not be able to find meaningful clusters due to overfitting or to uniform distances between data points.

Since BERT embeddings are token-based and not necessarily part of the same vector space as word vectors, they cannot be compared to each other, as was the case in *top2vec*, to determine the topic words. To solve this problem, Groostendorst developed a class-based TF-IDF method (c-TF-IDF) which is the last portion of the approach. Regular TF-IDF can only deliver relevant words for a given document. Therefore, it needs to be modified to deliver relevant terms for an entire topic or collection of documents.

This is achieved by merging all documents in a single category cluster into one big topic document. Then the TF-IDF score for terms in each topic document is calculated. This produces relative importance values for words in the document clusters. The most important words serve as a representation of the topic. Even though their frequency suggests an importance for the underlying collection of documents, it does not mean that said words form a coherent topic. Therefore, Maximal Marginal Relevance (MMR) is applied to determine the most coherent terms while reducing overlap or redundancy. It calculates the similarity of each topic word with the topic document while also considering the similarity between the topic words themselves (Carbonell & Goldstein, 1998).

In order to reduce the total number of topics, the least frequent topics are merged with their most similar pair, as was the case with top2vec. Their distance is measured by the distance between the feature vectors established with the c-TF-IDF method. This process can be repeated until the topics are clearly differentiable (Grootendorst, 2020).

## 2.3 TF-IDF

Term frequency-inverse document frequency (TF-IDF) is a calculation method developed to identify key terms for a single document in a collection of documents. The significance of a term is measured by its relative appearance (TF) in a document compared to the whole document corpus (IDF). If the number of mentions

of the term in the document is significantly higher than in the corpus, it is considered essential and defining to the given document. A normalization scheme for TF is applied in many cases due to the skew that might be introduced depending on the document length. For a term i in a document j, the frequency of the term is therefore divided by the frequency of all words in the given document $j$ $\left(maxOthers(i,j)\right)$.

$$TF(i,j) = freq(i,j)/maxOthers(i,j)$$

Terms that appear in all documents frequently should be accounted for with less significance because they carry low distinctive information. Therefore, the IDF is calculated using the logarithm of the number of all documents N and dividing it by the number of documents containing a specific term $i$ $(n(i))$.

$$IDF(i) = \frac{\log N}{n(i)}$$

TF and IDF are multiplied, which results in the final score for a given term in a document (Musto, 2010).

$$TF - IDF(i,j) = TF(i,j) * IDF(i)$$

# 3 Methodology

We conducted the research in the context of personalized radio broadcasting. The goal was to mimic a user's experience when interacting with a streaming app for radio content when selecting or specifying topics of interest, as many news and radio apps enable them to do. Based on this, the research process is divisible into three essential parts, which are the following:

1. Construction of Categories
2. Implementation as App-Prototype
3. Qualitative Interview

The first part deals with constructing the different topic models, which will be compared to answer the research question. Namely, the three category systems are *fully automatic, hybrid,* and *manual.* An in-depth explanation of each category system and details on its construction can be found in the next chapter. The second part is the construction of the app prototype, which was presented to participants. A breakdown of the implementation details and information on the architecture can be found in the chapter "Implementation." In the third part, the qualitative interview, participants were asked to interact with the prototype and describe their experience with the different topic models. A detailed description of the interview procedure can be found in the chapter "Design of the User Study". Finally, the results of the interview were evaluated and discussed in order to answer the research question.

## 3.1 Derivation of Categories

Three methods of deriving categories were compared to each other. The differentiating factor is the degree of manual effort that is necessary. We conducted our study in the context of a German nationwide public radio broadcaster and worked with a dataset of over 23 thousand documents that cover transcribed teaser texts of individual radio contributions published in an online media library. Furthermore, the broadcaster maintains a process where experts and archivists assign tags to the manuscripts. The sum of all tags is organized in a hierarchical taxonomy consisting of 12,236 entries. Finally, the broadcaster also provides a mobile app in which the user

can select preferred categories from a collection of 34 terms dealing with different topics, which were also selected by experts curating the media library.

**Manual Method (Baseline)**

The 34 categories provided in the app serve as a baseline and thus form the first method of deriving categories, the fully manual method. They have been developed by specialists working with the data and media library constantly.

Regarding attributes, they are non-hierarchical and cover multiple dimensions such as topics like "Germany" but also radio formats like "Interview".

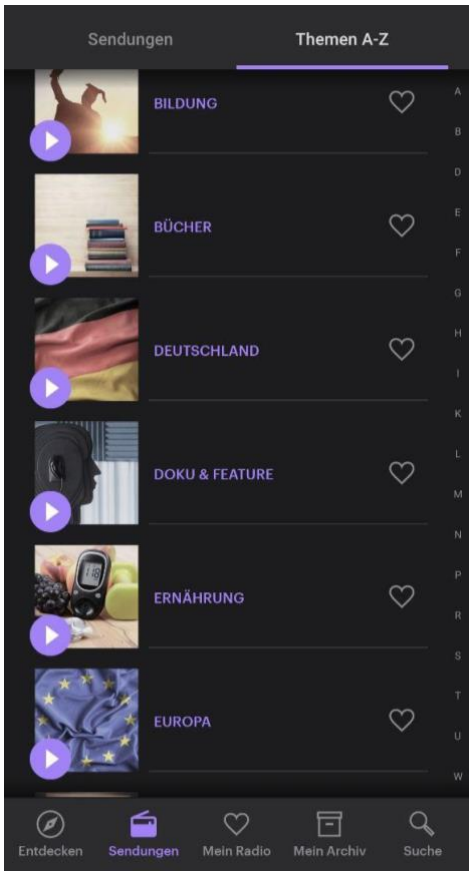


Figure 5: Radio-App Interface

**Automated Method (Unsupervised Topic-Modeling)**

The second method is the fully automated method. The aim of this approach was to eliminate manual intervention to a maximum. It resembles a bottom-up

approach as the models learn directly from the data and do not carry a priori assumptions.

The process for the automated method was primarily based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Wirth & Hipp, 2000). In general, the CRISP-DM framework provides guidance over the lifecycle of a data mining project. It is separated into six different phases with a general but not fixed sequence. The research is generally conducted sequentially along with the phases, but researchers may move back and forth between the phases depending on the findings during the research.

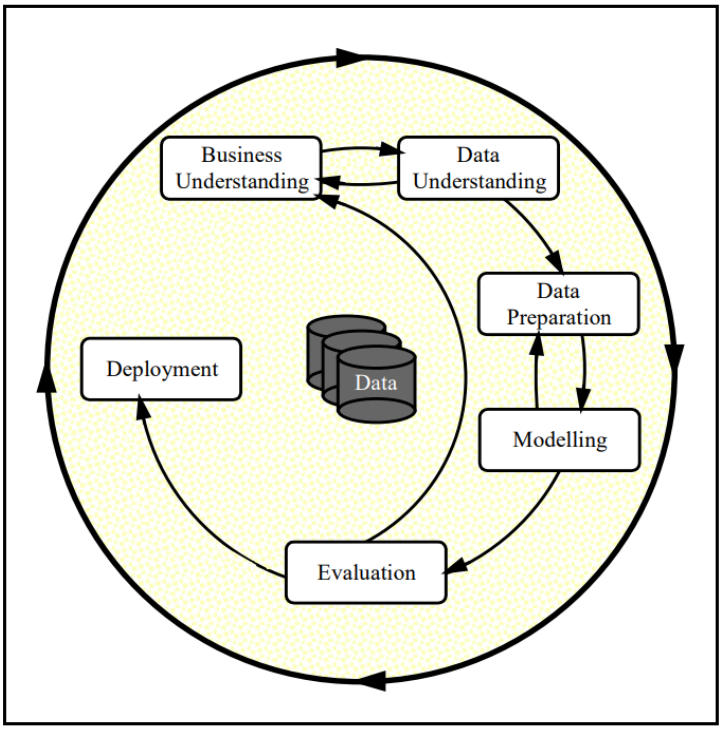


Figure 6: CRISP-DM Cycle (Wirth & Hipp, 2000)

In figure 6, the lifecycle process is illustrated. The outer circle symbolizes that the entire research process is repeatable indefinitely. In our case, that means that further research on automated topic models should be conducted in the future based on our findings. The cycle starts with the first phase, which is called *Business Understanding.* During this phase, information on the underlying problem as well as objective is gathered.

Furthermore, this process analyzes which data sources may provide informative data and can be considered for the further research process. Assessing the situation such as given inventory of resources, risks as well as cost and benefits is also part of this process. Finally, data mining goals and success criteria are determined so a project plan can be developed. The following phase is called *Data Understanding* which deals with the overall data gathering and exploration step. During this phase, data is gathered from one or multiple sources. Then it needs to be described and explored in order to determine its quality. In the *Data Preparation* phase, based on inclusion and exclusion criteria, data is selected. Next, the data is cleaned and formatted, and attributes are derived, depending on the requirements. The following phase is the *Modeling* phase. Here modeling techniques are selected. Depending on the case, multiple modelings can be taken into consideration. The models are then built, which includes parameter setting and their optimization. Finally, the model is assessed, and well as parameters revised. In the fifth, the *Evaluation* phase, the entire results are reviewed. This includes assessing the data, mining results, and models based on the criteria developed in the first phase. Additionally, the overall process is reviewed, and subsequent possible actions are to be determined so the final model can be deployed, which is the last phase. In the *Deployment* phase, the monitoring and maintenance plan for the final system needs to be established and documented.

Concerning our research, in the first phase, we established that a fully unsupervised topic-modeling approach was needed. Furthermore, it was clear that computing and time resources are limited and need to be considered. A German public radio broadcaster provided us with multiple datasets that were considered in the research process. The data mining goal was to develop a model which can generate coherent topics out of datasets in the German language. Phases two to five are explained in the following section. They were conducted in a non-sequential manner, going back and forth between the phases until an optimal model was found.

Multiple unsupervised topic-modeling approaches were tested to determine which delivers the best results and will be used in the study. Namely, the three algorithms are LDA, Top2Vec, and BERTopic. All of them were applied to the dataset provided by the radio broadcaster. Detailed information on the procedure such as the data, preprocessing, hyper-parameter optimization, and each model can be found in chapter 4, "Implementation".

It has to be noted that the algorithmic topic models do not provide single terms to describe topics but a collection of words from the actual document corpus, which, according to the model, are defining for the topic. Thus, some manual intervention was necessary to summarize the topics into one defining term. The specifics will be explained in the latter part.

Due to the unsupervised nature, objective evaluation and selection of models is challenging (Wallach et al., 2009). In contrast to supervised models, we do not possess "true" answers and thus can not compute objective metrics such as accuracy, precision, or recall. Therefore, we need to determine our own metrics specific to the application and data. In literature, various methods for evaluating topic models are presented, such as perplexity or held-out likelihood. Those methods are suitable for evaluating predictive models but fail to address the explanatory power of topic models that are of interest in this study (Chang et al., 2009). A more suitable evaluation method for the underlying study is direct human evaluation. Chang proposed two significant tasks in the human evaluation of topic models (Chang et al., 2009). The first one being *word intrusion* and the second *topic intrusion.* Word intrusion deals with the coherence of topics and the relationship between the terms given in each topic. For example, in a collection of words such as {*cat, dog, mouse, banana*} banana could be identified as an intruder as it does not fit in the category "animal" that the majority of terms belong to. Topic intrusion deals with the fit of each topic to its assigned document. This task is more concerned with classification accuracy and is therefore of less relevance to our study.

Based on this theory, combined with the specific requirement to develop a category system, these three major metrics were used to determine the best models.

- *Number of Topics*
- *Quality of Topics*
- *Number of unclassified documents*

The first metric is necessary to eliminate models which deliver too few topics. That would indicate that the model is not able to differentiate topics sufficiently. The minimal threshold number is 18 topics. We arrived at this number because it is the number of root nodes in the taxonomy, which describe the highest level of abstraction

of topics in the dataset. Minor deviations below the threshold number were tolerated depending on the scoring in the other metrics.

The second metric is comparable to the prior described word intrusion method. It aims to assess the overall quality of topics by analyzing the coherence and the overall quality and expressiveness of the topic words. Topic quality is the most crucial factor as it builds the basis of the categorization system. Each model is inspected for the following types of topics: topics that are coherent and can be clearly assigned to a single term, topics that show an indication but are ambiguous, and topics that show no sign of coherence and are incomprehensible. The larger the portion of topics belonging to the first category, the better the overall topic quality of the model is.

Finally, the number of unclassified documents acts as an auxiliary metric. Models that perform similarly on the prior two scores and have similar overall topic quality are also assessed by comparing the number of unclassified documents. Ideally, the model classifies as many documents as possible, which indicates that it managed to encompass large parts of the document corpus in the topics.

After selecting the best model and applying the topic modeling on the dataset, the collection of topic words needed to be distilled down to one defining term per topic in order to be able to be used in the prototype. The procedure was the following: the collection of topics produced by the model was given to four individual researchers. Each of them assessed the topics independently and assigned a summarizing keyword for the topic. Finally, after all, topics were distilled, the term was assigned if at least three researchers suggested the same term. For all other topics, the final term was decided over in a discussion.

**Hybrid Method**

The final method is a hybrid method aimed at combining elements from the manual and automated methods. In literature, multiple approaches for interactive topic modeling have been proposed. Hybrid approaches can help eliminate downsides from either approach, like incoherent topics developed by automated topic models. Furthermore, they enable the researcher to leverage personal knowledge in the field and introduce it to the unsupervised method (Hu et al., 2014). For many researchers, automated topic models are often a "take it or leave it" proposition. A hybrid approach

enables us to discard this notion and build on what the topic model produces in order to achieve a holistic result (Hu et al., 2014).

Building on this theory, our third approach combines automation and manual and bottom-up and top-down techniques by utilizing the taxonomy provided by the radio broadcast archivists to refine the output generated by the topic model. The goal was to introduce a top-down structure and context between the topics and a cleaner distribution over multiple subjects. Contrary to the prior two methods, we generated a hierarchical structure and arranged topics into main and subgroups. The main category, for example, might be "Sports" with subcategories such as "Soccer" and "Basketball".

We started by using the identical categories produced in the fully automated method by the best model. Next, for each category found by the algorithm, the occurrence or closest corresponding category from the taxonomy was noted. All terms and their corresponding root nodes are added to the category system. Furthermore, the taxonomy was examined for sibling categories that complement the collection. If the algorithm found the topics {Basketball, Soccer} and a sibling category would be {Tennis} it would be included as well. The remaining categories, which were either too granular, specific, or not represented in the taxonomy, were manually structured and aggregated by the researcher. Fitting topics covering the scope of the remaining terms and fitting into the already established categories were added to the system. By adding the terms and their corresponding root node, the hierarchy element is introduced with a total of two layers: parent and subcategories. So, in total, the final category system contains categories from the following sources:

1. direct matches between categories produced by the model and the taxonomy
2. parent categories of direct matches
3. fitting sibling categories of direct matches
4. additional categories constructed by the researcher

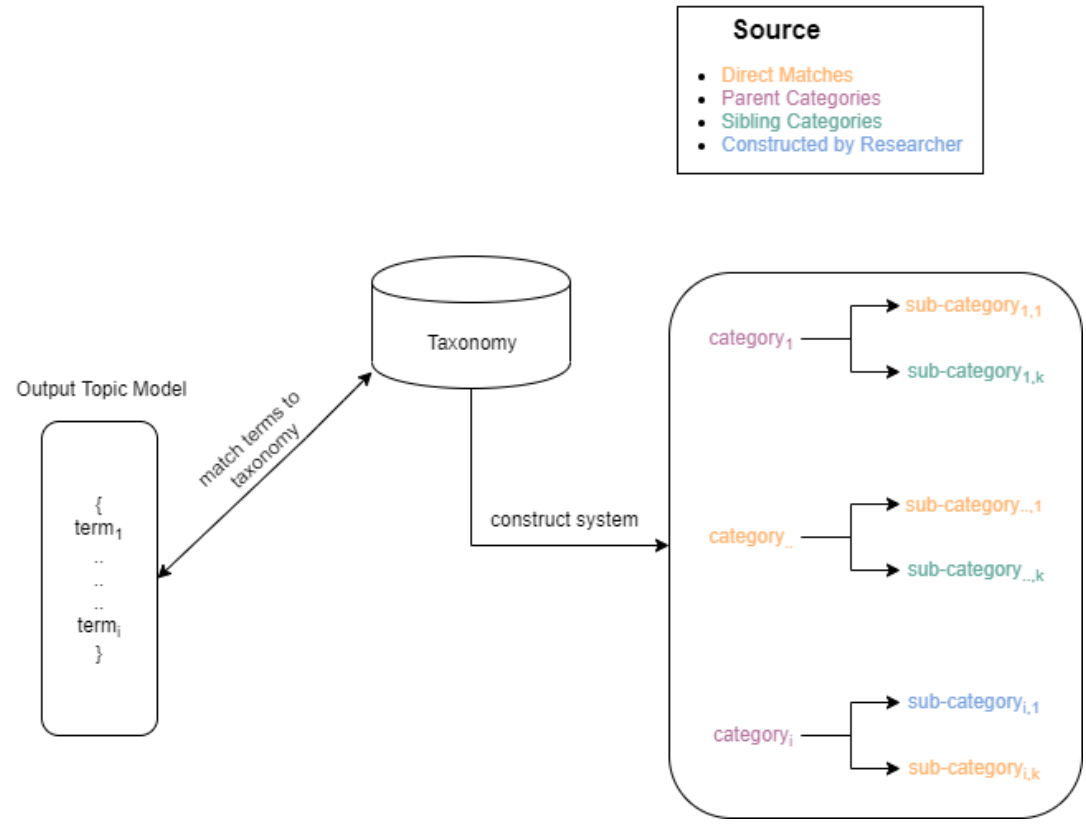


Figure 7: Hybrid Approach method

## 3.2 Evaluation Design

For the qualitative interview, the think-aloud method is used. The aim is to collect as much qualitative and in-depth data regarding the category systems as possible. This method is especially suited as it encourages the participant to describe any thoughts and ideas unrestrictedly while interacting with the prototype (Bruun & Stage, 2015).

On the spectrum of think-aloud methods, there exist methods with various degrees of influence by the interviewer. As described by Ericsson and Simon (Ericsson & Simon, 1993), the method in the traditional sense should be conducted without any intervention by the interviewer. Diverging from the traditional approach, there also exist newer methods like the *Speech Communication Think Aloud* developed by Boren and Ramey (Boren & Ramey, 2000). This variation of the method allows for a medium interception in the form of confirmations or repetition of the participant's last phrases to keep the conversation flow going. Finally, the method with the highest amount of

interviewer interception is the *Coaching Think Aloud* method. It allows for direct questions and guidance to specific areas in the interview (Hertzum et al., 2009).

In our interview, we decided to use the *Coaching Think Aloud* method. It offers an adequate balance of structure, focusing the participants' attention on the actual category systems and freedom for the participant to voice their opinions and thoughts unrestricted. The participants were asked to interact with the prototype as if they just downloaded the streaming app and were to "like" their desired categories. Meanwhile, they were supposed to think out loud and talk about their experience. The interviewer made clear that the focus of attention should lie on the actual categories and not other factors such as user interface design.

Furthermore, the meaning and functionality of the user interface were explained to the participant in case some elements were not understood. This procedure was conducted for all three category systems. Finally, the participants were asked to voice any final thoughts and what was perceived well and what was not about each category system. All interviews were transcribed for further analysis.

Ten people between the ages of 20 and 47 were interviewed during June 2021 in the course of the study. To prevent a gender bias, the selected participant consisted of 5 males and five females. We decided to interview people in this age range as they would be the typical users for a mobile application in the given radio context. Younger demographics are typically not radio listeners yet, while older demographics tend to use a more traditional medium than a smartphone. An emphasis was placed on interviewing people from diverse backgrounds to gather opinions and information independent of background or profession. The participants' occupations covered multiple branches such as university students and office workers or manual professions such as carpentry. Roughly 50% of the participants were active radio listeners, while the other 50% tended to listen more to other media such as music streaming, podcasts, etc. but still showed an interest in radio content if it became available in a more personalized/on-demand manner. All interviews were recorded and transcribed to enable the construction of the coding frame and analysis as described below.

**Data Analysis**

The transcriptions were analyzed via qualitative content analysis. This data analysis method deals with the structuring of data in a bottom-up approach. The basis for the data extraction is what is called a "coding frame". It is constructed by assigning fitting categories to all parts of the underlying data. Depending on the aim and if the analysis of themes is sufficient, the frame can even be the result itself (Mayring, 2014, p. 84). Working with a coding frame is beneficial as it introduces structure and a system into the exploratory analysis.

Furthermore, it is transparent and enables other researchers to replicate and build on findings. Due to the exploratory nature of our study, we opted not to work with a priori assigned categories for the coding frame. This would limit the potential findings and restrict the latent space in which we are moving. In 2000 Mayring proposed a method for manual topic modeling (Mayring, 2000). In summary, it describes a bottom-up approach, where categories are deducted from the data directly in contrast to deriving ones from theory. The method is structured in six major steps (Figure 8). It starts by establishing a primary criterion for the selection of relevant parts in the material. This criterion is the direct link to our research question that can be formulated as *aspects that impact user experience with the category systems*. By keeping the criterion on a high meta-level, we enable the construction of the coding frame via a bottom-up approach.

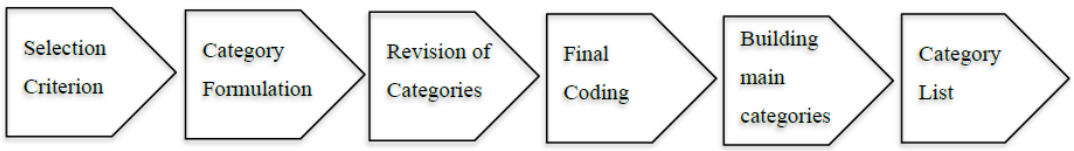


Figure 8: Category Development (Mayring, 2014)

In the second step, each transcription is worked through sentence by sentence. Every time the criterion is met and new aspects are named, they are noted as potential categories for the coding frame. After working through all transcripts, the categories are revised, summarised in case they are highly similar or redundant, and structured. Next, a frequency analysis is conducted; aspects that are named in higher quantities are considered to be more critical than rare aspects. A table is constructed, which

comprises the coding frame with all aspects. Finally, the transcripts are analyzed with the coding frame in mind, and the table is filled with the opinions of the users regarding each aspect.

# 4 Implementation

In the following chapter, the implementation of the topic models and the prototype will be explained. It starts with the description of the raw data and its preprocessing. Next, the data modeling process is elaborated. We explain all models that were tested and all relevant parameters. Finally, the UI developed for the user study is shown.

## 4.1 Datasets and Preprocessing

We had to decide between two different datasets to work with. Both of them are in the German language and cover the contents of a nationwide public radio broadcaster. The first one was a collection of 63,165 documents which are transcripts of existing radio contributions. Each document contains at least 100 words and covers different topics as well as formats. Some texts, for example, are structured as interviews, while others are continuous texts. All of them contain a large amount of metadata. The other dataset contains around 23 thousand documents composed of short descriptive text snippets covering one radio contribution each. They cover various topics as well but are all in a uniform format. Each document contains a short text of approximately 50 words and a title. Furthermore, the first dataset covers older radio contributions (around 2015), while the second one only covers newer contributions (2020-2021) and is updated regularly. The second dataset is considerably better in terms of data quality due to the higher uniformity of each document and cleaner formatting and structuring.

The first dataset was provided to us directly in the form of individual text files. The second dataset was fetched over an API via a python script into a .csv file. In each row, one element was represented with various features. We cleaned the data by dropping all rows with missing features. The resulting data frame contained the following features (Table 1):

Table 1: Data Features

| **Attribute** | **Data type** | **Description** | **Example** |
|---|---|---|---|
| audio_id | Integer | A unique number that serves as a primary key. | 921398 |

| | | | |
|---|---|---|---|
| title | String | Represents the title of the radio contribution. | ‘Woher kommt der grüne Wasserstoff - Küste oder Wüste?’ |
| authors | String | Represents the name(s) of the author(s). | ‘Grotelüschen, Frank’ |
| teaser | String | A brief description (approx. 50 words) of the contribution's content. | ‘Grüner, regenerativ erzeugter Wasserstoff – das ist die Vision, in die Milliarden investiert werden sollen. Aber wie? Produktion in Deutschland, zum Beispiel mit großen Windparks an der Küste? Oder Import, etwa aus sonnenreichen Regionen wie der Sahara? Gebraucht werden jedenfalls riesige Mengen.’ |
| broadcast | String | The name of the broadcast the contribution belongs to. | ‘Wissenschaft im Brennpunkt’ |
| duration | Integer | The total duration of the audio in seconds. | 1789.0 |
| date | Datetime | A timestamp when the contribution was published. (XX/2020 - XX/2021) | 2021-05-09 00:00:00 |
| audio_path | String | URL of the audio resource. | path to audio file (no example for data security reasons) |
| image_small | String | URL of the small image resource. | path to image file (no example for data security reasons) |
| image_large | String | URL of the large image | path to image file |

| | | | |
|---|---|---|---|
| | | resource. | (no example for data security reasons) |

All attributes were relevant in the context of the prototype, but for the topic modeling relating to only the "teaser" and "title" columns were needed.

Both sets were preprocessed in the same manner. The goal of preprocessing is to convert documents into an organized structure that is analyzable and predictable by a Natural Language Processing algorithm. A model's performance is highly reliant on the result of the preprocessing; therefore, one needs to be especially considerate during this step. The exact steps are highly dependent on the domain as well as the algorithms that are used. This leads to the preprocessing step being partially a trial and error process. In our study, the following preprocessing steps were taken:

1. **Lemmatizing:** All words are transformed into their canonical form, their word root, or lemma. For example, the word "running" would be transformed into "run" or "apples" to "apple". This step helps the algorithms to identify the same words. Otherwise, the same word in different forms might be interpreted as different words.
2. **Stop-word Removal:** English, as well as German stop-words, are removed. Stop-words are words such as "this, is, at, the, etc.". Those words do not add significant meaning to the sentences and only add noise to the models.
3. **Lowercasing:** All words are cast to lower-case. Without this procedure, words that are the same but capitalized (e.g., at the beginning of the sentence) and non-capitalized in another place would not be recognized as the same word.
4. **Remove all symbols except letters:** The documents contain many symbols such as numbers, punctuations, and line breaks. Those added noise and were therefore removed.
5. **Tokenization:** Each text document is split into single terms using the state of the art library "nltk" (*Natural Language Toolkit — NLTK 3.5 documentation*, 2021).

As demonstrated by Camacho-Callodos et al., the performance of models is highly variable depending on the preprocessing decisions, and in the case of text classification based on neural networks, decisions like lemmatizing might even reduce performance (Camacho-Collados & Pilehvar, 2018). Therefore, a brief test run was conducted for the models (Top2Vec and BERTopic), which may be negatively impacted by extensive preprocessing. It was evident that the preprocessed data sets produced way more comprehensible topics. Thus, we opted to conduct all above listed preprocessing steps for all models.

Additionally, to the basic preprocessing steps, two further steps were conducted specific to the given research problem. Category terms only consist of nouns. Therefore, a consideration to improve the model's performance was to filter the datasets for nouns only. Additionally, it was not clear if adding the title of each document to its text corpus was beneficial for performance or not. Based on those considerations, we ended up with four variations of the preprocessed dataset being the following:

- All words and no title
- All words with title
- Nouns only and no title
- Nouns only with title

Initially, all three approaches (LDA, BERTopic, Doc2Vec) were tested to determine the best one on the datasets with all words, including the title. The best of the three approaches was then tested on all of the four datasets to extract the best model.

## 4.2 Topic Models

Hyperparameter optimization is of the highest importance. The choice of which hyperparameters to optimize is highly dependent on the underlying algorithm. Therefore, the optimization process will be explained for each model.

The only hyperparameter that was optimized on LDA was the number of topics. Even though a second parameter that may be optimized is learning decay, which impacts the learning rate, we found that it did not have any noticeable impact on the model's output in the test runs. Therefore, working with one hyperparameter

only reduced computational efforts significantly and was a tradeoff we were willing to make considering the large number of documents and computation times for LDA.

- Parameter : *no_topics [2;120]*
- Metric: *coherence_score*

The tested range for the number of topics was from 2 to 120 with a step size of 2. The model was optimized on the coherence score. It measures the grade of semantic similarity between the main terms in the topic. The aim is to distinguish between semantically interpretable (high coherence) topics and topics that are a product of mere statistical inference (low coherence). Furthermore, an optimal coherence score indicates enclosed topics that are semantically coherent but do not have high overlap with other topics (Röder et al., 2015).

The coherence model presented by Röder et al. is implemented in a python library called gensim, which was used in the study. It consists of a four-stage pipeline containing Segmentation, Confirmation Measure, Probability Estimation, and Aggregation. In the first step, word sets (topics) are segregated into word pairs, where every word is paired with every other word in the set. In the second step, the confirmation or coherence between a given pair is estimated. Different ways of word probabilities can be used, depending on the context of the application, which forms the third step. Finally, the scores are aggregated to form a summarizing coherence score for all word sets (Röder et al., 2015).

In figure 9, we can see that the highest coherence score *c* was reached at 74 topics with a score of 0.53.

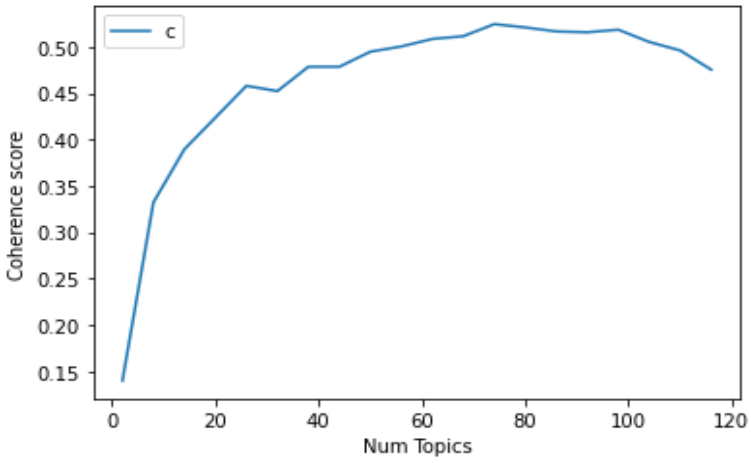


Figure 9: Coherence LDA

After optimizing the model, the topics were visualized. In figure 10, it is evident that the optimization method also produced topics with few overlappings and good distribution across the principal components:

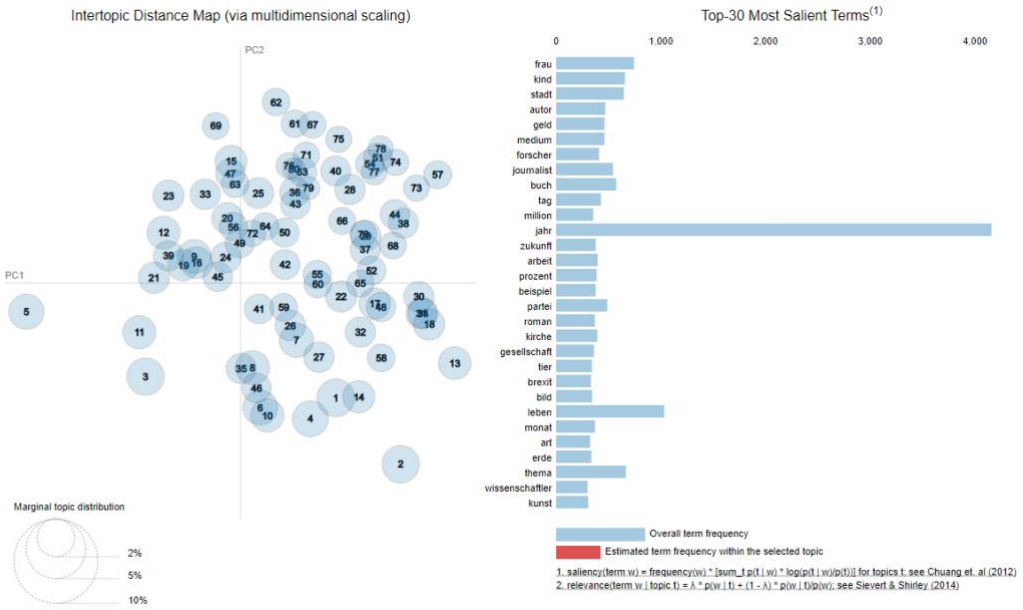


Figure 10: LDA Intertopic Distance Map

Manual inspection of the actual topics showed that even though metrics showed an even distribution of semantics, the topics themselves contained much noise and were not necessarily coherent for a human interpreter. Most of the topics (87%) showed a thematic indication but could not be pinpointed to a specific topic. Especially

the largest topics containing the largest amounts of documents did not show any semantic coherence but seemed like a collection of "frequent" terms. Therefore, other models had to be tested to determine if results could be improved upon, as LDA did not produce satisfying results.

For Top2Vec, there are two hyperparameters which are the following: *speed* and *workers*. '*Workers'* determines the number of worker threads used to train the model. This parameter depends on the machine that the training takes place on. In our case, we were able to use a maximum of 8 workers. *Speed* sets the model's training speed; there are three options: *fast-learn*, *learn*, and *deep-learn*. Due to the moderate size of our dataset, we opted to work with the *deep-learn* option because the training time was still manageable, and more thorough training produces the best results.

The algorithm produced a total of 136 topics after execution. In order to analyze all topics and gain insights into the quality of the results, all topics were displayed as word clouds which is a feature of the Top2Vec python library (Angelov, 2020/2020). The word clouds enabled us to gain fast insights into the coherence and quality of topics.

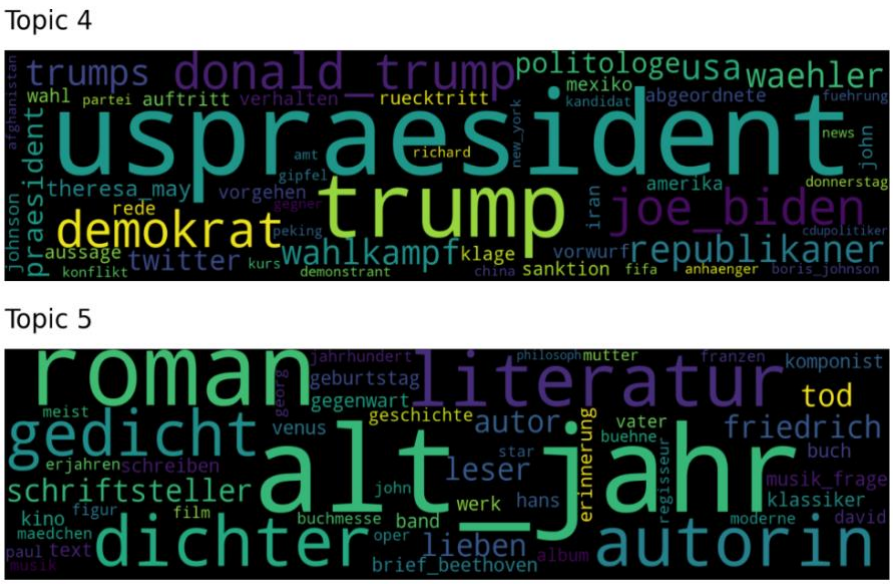


Figure 11: Top2Vec Wordcloud

Out of the 136 topics produced by Top2Vec, 47% of topics were coherent and could be assigned with a general topic term, 23% were vague and showed a general

direction of a topic but contained a larger amount of noise, and 30% were incoherent and did not show semantic context. A potential reason for those mediocre results (this applies to LDA as well) might have been the short length of documents. The library's author states that Doc2Vec performs best on large datasets with unique vocabulary (Angelov, 2020/2020). Machine learning models, in general, perform better the larger and higher quality the training data is. Therefore, we suspected that we might achieve better results by utilizing pre-trained models which were trained on extensive datasets compared to models like LDA and Top2Vec, which had to be trained on our data from scratch. For this reason, our last test was conducted by utilizing pre-trained sentence encoders.

For BERTopic, there are multiple options to optimize the performance. The first choice is to determine which pre-trained model to use. We decided to test five different models which were trained on different datasets and have a diverse collection of base models to cover a wide range of possibilities. Each model has different strengths and weaknesses, and we were not sure what would be the most effective one on our particular dataset. Thus, we decided to select multiple models covering different base models and training datasets. In table 2 all tested models with their corresponding pooling method, the base model, and the training dataset can be found.

Table 2: Pretrained Models

| **Model Name** | **Base Model** | **Pooling** | **Training Data** |
|---|---|---|---|
| distilbert-nli-base-mean-tokens | distilbert-base | Mean Pooling for CLS tokens | NLI |
| stsb-roberta-large | roberta-large | Mean Pooling | NLI + STSb |
| stsb-distilbert-base | distilbert-base-uncased | Mean Pooling | NLI + STSb |
| paraphrase-xlm-r-multilingual-v1 | XLM-R | Mean Pooling | Paraphrase Data |
| nli-bert-base-cls-pooling | bert-base-uncased | CLS Token | NLI |

All five models were tested on the dataset containing all words, including the title. They were compared for the number of unassigned documents as well as the quality of topics. The topic quality is determined by the number of topics that can be summarized into a single topic term. So, if 40 out of 100 documents can be clearly identified as one coherent topic, the topic quality would result in 0.4. In table 3, the results can be seen.

Table 3: Comparison Model Performance

| **Name** | **Topic Quality** | **No. Unassigned Docs** |
|---|---|---|
| distilbert-nli-base-mean-tokens | 0.74 | 5676 |
| stsb-roberta-large | 0.75 | 5702 |
| stsb-distilbert-base | 0.7 | 6743 |
| paraphrase-xlm-r-multilingual-v1 | 0.65 | 6532 |
| nli-bert-base-cls-pooling | 0.67 | 6923 |

All five tested sentence encoders beat our prior approaches (Top2Vec, LDA) regarding topic quality. Furthermore, it is notable that all of the sentence-encoders showed significantly shorter training times (3-10 minutes) compared to LDA (>1 hour) and Top2Vec (>30 minutes), even though the models are larger and show higher complexity. The reason for this is that LDA and Top2Vec needed to be trained on the CPU while the sentence encoders support GPU computation which parallelizes most of the computing tasks (*Pretrained Models — Sentence-Transformers documentation*, 2021).

When comparing the sentence encoders, we can see that the best two models were *distilbert-nli-base-mean-tokens* and *stsb-roberta-large*. Their topic quality was

the highest, and the number of unassigned documents the lowest. Even though *stsb-roberta-large* had a 1% better topic quality, we decided to work with *distilbert-nli-base-mean-tokens* for further optimization since the training time was significantly shorter and the model is smaller even though it provided comparable results.

Other hyperparameters relevant for BERTopic besides the embedding model are the minimum topic sizes (*min_topic_size*), as well as the number of topics which is the number of clusters to be determined (*nr_topics*). The author of the python package himself recommends choosing a minimum topic size that is not too restrictive but high enough to weed out topics that contain single documents. After a trial and error process, we found that the ideal minimum topic size was 20 documents per topic. For the number of topics, the author implemented an automatic option (*nr_topics = "auto"*) which merges all topics whose c-TF-IDF vectors show a cosine similarity of over 0.9. Thus, smaller, similar topics are merged into bigger ones, and the overall number of redundant topics is minimized.

Finally, the model was applied to the four preprocessed dataset versions to determine if topic quality could be improved further. We developed the theory that Topic Quality might increase by filtering the document for nouns only because a lot of the incoherence and ambiguity stemmed from verbs and adjectives in the topics. The results from the application on the alternative datasets for the best BERTopic model (distilbert-nli-base-mean-tokens, nr_topic=’auto', min_topic_size='20') can be found below (Table 4).

Table 4: Comparison Datasets

| Dataset | No. Topics | Topic Quality | No. Unassigned Docs |
|---|---|---|---|
| All without Title | 106 | 0.73 | 6534 |
| All with Titel | 105 | 0.74 | 5443 |
| Nouns without Title | 107 | 0.83 | 5581 |
| Nouns with Title | 101 | 0.86 | 3743 |

The results show that the number of topics calculated by the model is similar for all datasets, and the difference is negligible. However, in performance terms, working with nouns only and including the title leads to a tremendous increase. The topic quality increased as expected (+ 10-12%) for both variants (with and without title) after reducing the dataset containing only nouns. The impact was comparable regarding the number of unassigned documents, but for the documents containing the title, the impact was more significant than for the documents without the title (-30% and -14%, respectively). Adding the title only showed a slight increase in topic quality (+1-3%) but a significant decrease in the number of unassigned documents (-17% and -33%). A potential explanation for this observation might be that reducing the dataset to nouns clears up the documents and, ultimately, the collection of words in each topic, making them more coherent and comprehensive. On the other hand, adding the title does not reduce noise but only adds more informational content in a distilled way (as the title is a summary of the document's content) to each document, enabling for more precise classification and assignment of it to the given topics.

Based on those results, the final model which was used for the data-driven categories was a BERTopic model with the following hyperparameters:

- embedding_model=’distilbert-nli-base-mean-tokens’
- nr_topic=’auto’
- min_topic_size=’20’

It was trained on the preprocessed dataset consisting of nouns only and containing the title.

## 4.3 Prototype

The prototype developed in this study is modeled after a classic interface, as one might find it radio streaming application. It simulates a standard user experience when interacting with the app and selecting preferred categories or topics after installation.

For the frontend, Angular as a framework was used. Its component-based architecture enables a modular, encapsulated logic for future expansion because the prototype was used in a parallel study. Furthermore, Angular is platform agnostic

which is a benefit when interviewing multiple people who own different devices. In the following, only the part of the prototype which is relevant for the study will be explained.

The view presented to the user features all categories of a given category system in a list. It encompasses two major functions accessible to the user:

- like function
- search bar

Those two functions can be found in many applications with a category selection functionality and aim to provide the user with a familiar experience. Design-wise, the prototype is kept to a minimum and makes use of Google's Material Design library, which is the standard for Angular applications (Angular Components, 2021). The use of a design library also ensures that best practices for user interface design are implemented, so the study participant is not distracted by design choices and focuses on the interaction and evaluation of the categories.

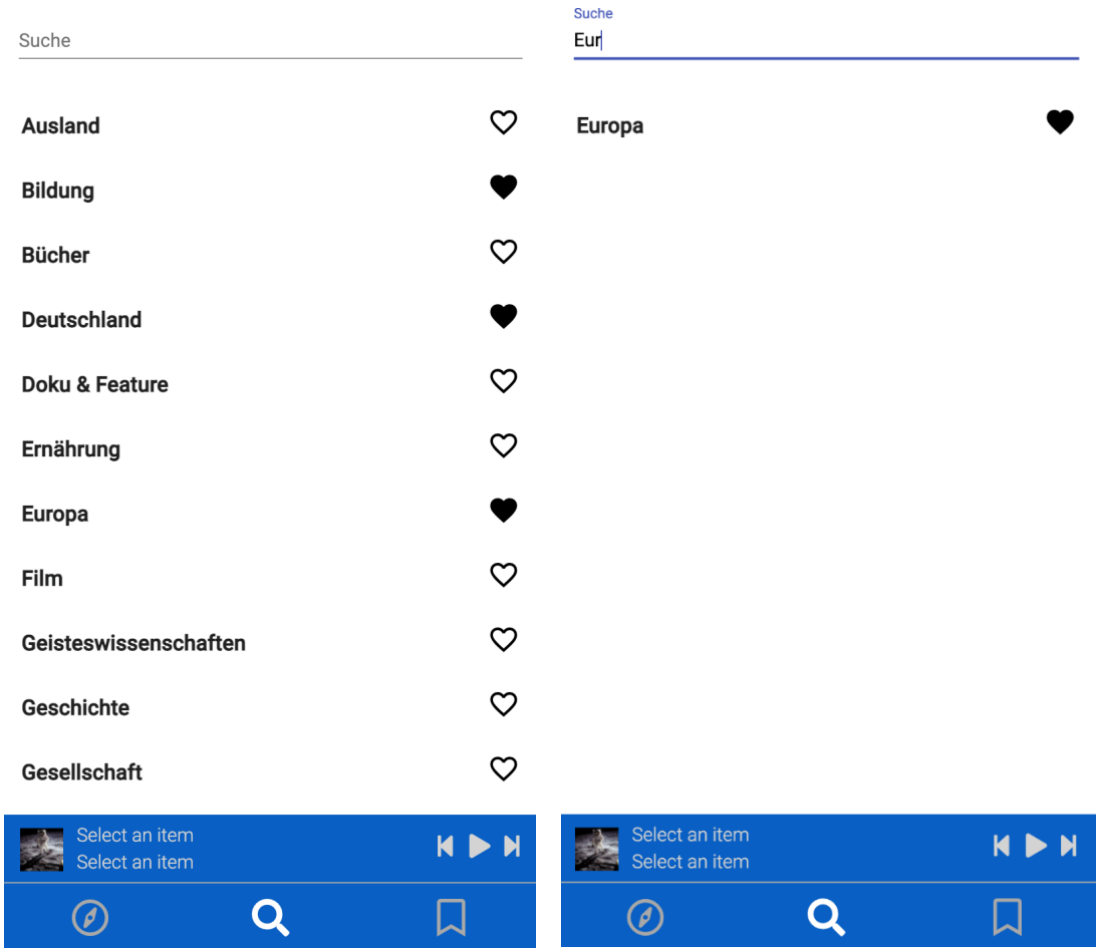


Figure 12: UI Like and Search Function

To switch between the category systems, a hidden feature for the interviewer was implemented. After typing "devtools" into the search bar, three buttons appear which select between the expert, data, or hybrid categories.

The hierarchical element in the hybrid category system is visualized by displaying the names of the parent categories in bold font while the subcategories are displayed in regular font below them (Figure 13).

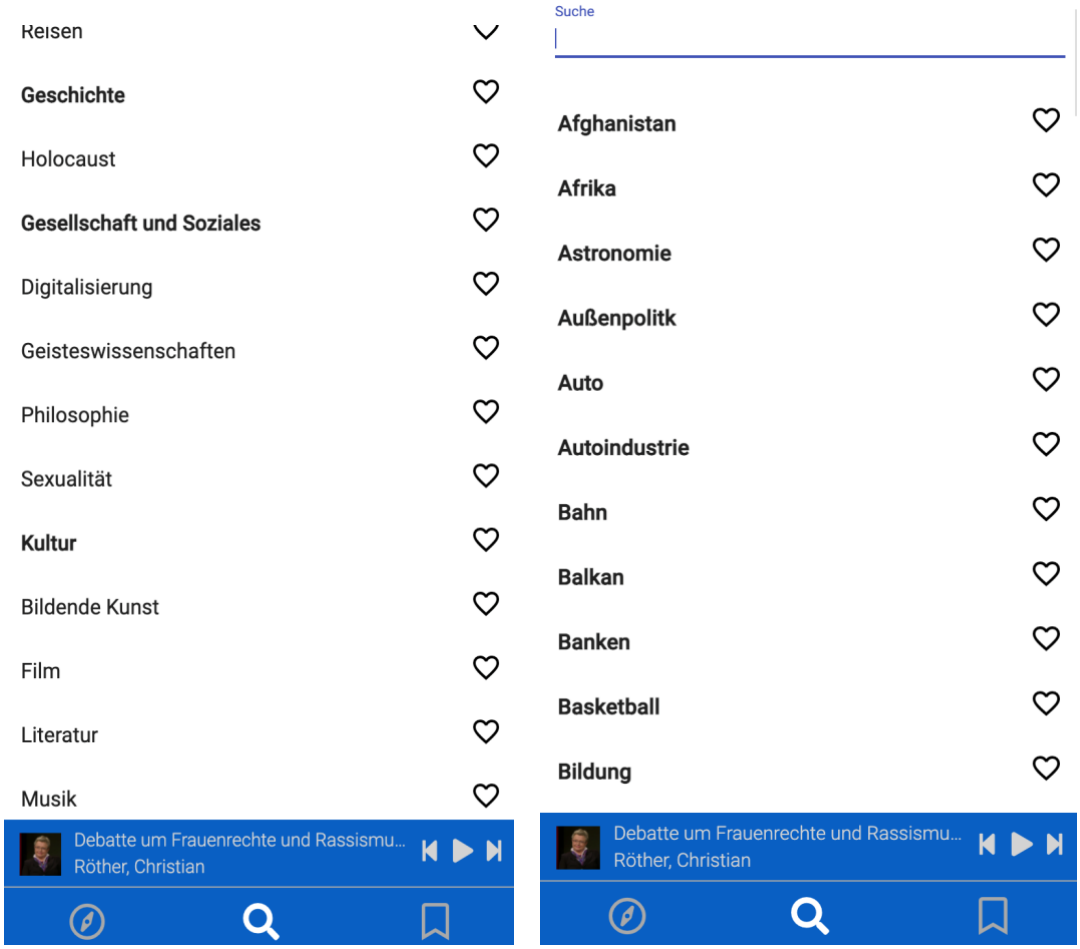


Figure 13: UI Hierarchy

# 5 Evaluation

In the following chapter, the results from the analysis of the user study will be presented. Ten participants interacted with the three different category systems produced by three different approaches, namely the (1) manual, (2) automatic, and (3) hybrid method.

A list of all categories to its corresponding system can be found in Appendix 1. Because the categories were directly translated from German, they are not in alphabetical order, as was the case in the original systems. The first system consists of 37 topics. There is no differentiation between different levels of abstraction, and all categories are displayed on the same hierarchical level. Furthermore, it features categories describing content like "History" and formats like "Radio Play". Overall, the categories cover higher levels of abstraction and do not cover specific themes. The second system covers 93 categories. Just as in the first system, all categories are on the same hierarchical level but cover different levels of abstraction. Formats and content categories are mixed as well.

Contrary to the first system, the second system does feature not only general, high-level categories like "Politics" but also concrete topics like "Black Lives Matter", thus covering multiple levels of abstraction ranging from high level to specific. The third system is hierarchical and features 68 categories divided into 17 parent and 51 child categories. Like the second system, it covers multiple levels of abstraction, but they are organized into two hierarchical layers. Furthermore, formats and categories relating to current topics are assigned to separate parent categories instead of mixed with other categories.

As suggested in the literature, the results will be presented by illustrating the coding frame resulting from the interviews' analysis (Table 5). This includes displaying all the final codes and their respective meanings and analyses. The final frame contains 17 categories, with each category informing how many participants mentioned it and whether they voiced a positive or negative opinion. In the following, all categories according to the coding frame are presented with an explanation.

**Number of Topics.** All participants voiced their opinion regarding the overall number of topics in the category systems. As mentioned above, the topic numbers were the following:

- *low*: system 1 with 37 topics
- *medium*: system 3 with 68 topics
- *high*: system 2 with 93 topics

For the first system, 8 participants said they felt the number of categories in system 1 was too low. They preferred a higher number to have a greater selection to choose from. Two participants were happy with the number of categories with the comment that too many categories may be "overwhelming" and they prefer fewer categories to select from in general. For the second system, 8 participants took a stand with mixed opinions. 4 participants disliked the high number of categories while the other 4 liked the high number. For the final system, only four people referred to the number of categories, all of them said that the number was "just right".

**Browsing Behavior.** The browsing behavior refers to how the users decide which categories to choose from. 7 out of 10 people said that they use the categories as inspiration and scroll through them to decide what to "like". The other three people mentioned that they are aware beforehand of what they are going to like and try to find their interest directly in contrast to being affected or inspired by the categories.

Out of all interviews, the search bar was only used by one participant briefly. We made sure that all participants were aware of the bar and how to use it. Therefore, not using it was an active decision. After asking for an explanation for why they did not use the search function, five people answered that the number of categories was too small to require the use of a dedicated search bar in all systems. Furthermore, the seven people that used the categories as inspiration did not know what to search for, nonetheless.

**Alphabetical Sorting.** All participants commented on the alphabetical sorting of categories. Seven people disliked the alphabetical sorting in systems 1 and 2 while 3 voiced positive opinions. After further inquiry, the people who disliked it elaborated that it was not the alphabetical sorting itself that led to the disliking but that it was the only form of organization in systems 1 and 2. If the systems were sorted hierarchically as in system three or grouped by themes, the alphabetical sorting would be fine. The

reason for this was that by sorting alphabetically only, thematic topics are distributed across the list. Cultural topics like "Music" and "Theatre" are mixed with "Politics" in-between, for example, instead of being displayed next to each other.

**Level of Abstraction.** This category refers to the level of abstraction of the given topics in the system. Topics range from broad-ranging categories like "Politics" down to highly specific topics like "Refugee Crisis". In regard to the systems, system 1 covered high levels of abstraction for the most part, while system 2 covered all levels of abstraction. System 3 also covered all levels of abstraction but featured less specific topics than system 2. 7 participants voiced their opinion regarding the low specificity of topics in system 1. They criticized the high level of abstraction and said that they felt like they could not specify their interests accurately enough. For systems 2 and 3, 8 people commented that they like the division into specific topics and events, especially the division in various sports categories, as they might be interested in news regarding one specific sport only but not regarding others.

**Hierarchy.** The hierarchical aspect was only featured in the third category system. In systems 1 and 2, all topics are displayed alphabetically on the same level, and there is no differentiation between parent and child categories. In 3, however, the categories were manually organized parent categories (bold font) and subcategories (regular font). All 10 participants noted that they experienced the hierarchical aspect very positively.

Contrary to the other two systems, five users mentioned that it did not feel "cluttered" but "well organized". Furthermore, the grouping assisted greatly during the navigation through the system as the users were able to skim over the parent categories to get a general overview of the topics and then dive further into categories of interest. All users did not only comment positively on the hierarchical aspect in system three but also comment negatively on the lack of hierarchy or grouping in systems 1 and 2. According to three participants, the lack of arrangement in groups combined with the mere alphabetical sorting introduced a feeling of "randomness and chaos" which was challenging to navigate. Instead of skimming over all categories and deciding which to examine further, they were "forced to read through every single category to know what is available to like".

**Content Arrangement.** This category includes all topics regarding comments that were made, referencing, and giving feedback regarding the content and its arrangement. All 10 participants commented on the mixture of formats like "Docu & Feature" and topics regarding specific content like "History" on the same level, as is the case in systems 1 and 2, in a highly negative manner. According to them, the structure was confusing as a certain format might contain all sorts of topics; thus, the presentation was not logical. The same participants commented positively on the variation in system 3, where all formats were grouped under one category. Nonetheless, some users noted that the inclusion of formats in the view, in general, is questionable as they felt out of place, and a separate view for formats might be more fitting as they have got nothing to do with topic and interest selection.

Furthermore, comments were made on the inclusion of time-relevant topics in the categorization system. 3 out of 4 people commented on topics like "Brexit" and "Corona-Virus" with the concern that topics like this quickly become outdated and may be too specific for the purpose of a general category view. On the other hand, one person noted that they liked the inclusion of those topics as they found high interest in certain ones and would be happy to listen to radio contributions revolving around this topic specifically. Nonetheless, they also mentioned that those time-relevant topics should be reworked regularly. All four people liked the separation of time-relevant topics in their own group, as was the case in system 3.

**Overall Preference.** At the end of each interview, the participants were asked which of all systems they liked the best and for what reasons. All 10 participants said that they liked system three the best. All of them said that the major factor that led to this decision was the hierarchical aspect and the advanced arrangement and organization of the categories. The other two systems felt "random" and difficult to navigate through. 4 out of the 10 people also said that they like system 2 as well for the selection and fine differentiation of categories. In their opinion, the ideal system would combine the fine abstraction from system 2 with the organization from system 3. In regard to system 1, all participants said that it felt too broad and unorganized.

Table 5: Coding Frame Evaluation

| Main Category | Talking Point | Total Statements | Positive | Negative |
|---|---|---|---|---|
| Number of Topics | Number of Topics in 1 | 10 | 8 | 2 |
| | Number of Topics in 2 | 8 | 4 | 4 |
| | Number of Topics in 3 | 4 | 4 | 0 |
| Browsing Behavior | Inspiration through Browsing | 10 | 7 | 3 |
| | Used Search Bar | 10 | 1 | 9 |
| Alphabetical Sorting | Alphabetical Sorting in 1 and 2 | 10 | 3 | 7 |
| | Alphabetical Sorting in 3 after grouping | 10 | 10 | 0 |
| Level of Abstraction | Abstraction in 1 | 7 | 0 | 7 |
| | Abstraction in 2 and 3 | 8 | 8 | 0 |
| Hierarchy | Hierarchy in 3 | 10 | 10 | 0 |
| | Lack of Hierarchy in 1 and 2 | 10 | 0 | 10 |
| Content Arrangement | Mix of Format and Topics | 10 | 0 | 10 |
| | Inclusion of Time Relevant Topics | 4 | 1 | 3 |
| | Separation of Time Relevant Topics in 3 | 4 | 4 | 0 |
| Overall Preference | Likes 1 | 10 | 0 | 10 |
| | Likes 2 | 10 | 4 | 6 |
| | Likes 3 | 10 | 10 | 0 |

# 6 Discussion and Limitations

## 6.1 Category Systems

With BERTopic, we found a topic modeling approach which produces state of the art topics and beat industry standards like LDA and Top2Vec in our use-case. Interestingly, the dataset that was preprocessed the most, performed best, even though literature suggests that sentence encoders should be used on datasets with little preprocessing as many contexts get lost. Furthermore, eliminating all words except nouns had the most profound impact on the coherence and clarity of topics. An explanation for this phenomenon might be that even though nuances and details get lost in the course of strong preprocessing, they are not as crucial in the context of topic modeling as topics are usually described by nouns, and only the essence of the document needs to be found out. In traditional applications of sentence encoders such as question and answer applications or sentence comprehension and completion, this kind of preprocessing would most likely result in worse results.

When deciding on a topic modeling approach, computing resources also need to be considered. In our case, we had a relatively powerful graphics processing unit (GeForce RTX2070 Super) available, which made the training of neural networks such as the sentence encoders feasible due to most libraries supporting GPU accelerated training through CUDA support (*CUDA Toolkit*, 2013). Training the same models on a regular CPU would increase training times exponentially, especially during hyperparameter optimization, where each training cycle would last upwards of 10 hours instead of a few minutes. Regarding further research, due to working with neural networks, the results and numbers we achieved might not be reproducible exactly as every rerun yields slightly different numbers.

Considering the research question, our primary focus was comparing the manual, data-driven, and hybrid topic modeling methods. Therefore, due to the given time and resource constraints, we could not focus solely on optimizing the best possible topic modeling approach. Further research should be conducted in this area, expanding on the optimization of the approach, working with different and larger datasets. Especially with the general approach BERTopic and Top2Vec are based on, modifications can be conducted at every step of the process. This includes working with custom embedding models and special fine-tuning for pre-trained sentence

encoders as we only compared different pre-trained models. For the dimensionality reduction with UMAP and clustering with HDBSCAN, we used recommended default parameters as well; thus, optimization in these areas can occur. Besides optimizing, further experiments with different clustering algorithms might also be conducted, which potentially might be more suitable for the given data.

To validate the generalisability of our findings, research on data in different languages besides German needs to be conducted. Additionally, testing the classification accuracy of the data model would be a point of interest. We only examined the topics themselves but did not check how accurate the actual classification of the model regarding each topic is.

As for the automated method, the output topics consist of word collections describing the topics. Thus, we had to assign the final topic term manually. To adjust for human bias, this process was conducted by three researchers independently. Ideally, this step should be eliminated in order for the method to become genuinely automated and objective.

The construction of the third system, the hybrid method, was conducted in a comprehensible and reproducible way as we laid out a framework on how each potential term was decided on. Generally, the process will look similar for most cases of manual rearrangement and cleanup of outputs by automated topic models (working with some taxonomy combined with subjective modifications by experts/researchers). Nonetheless, the resulting final categories may potentially be vastly different as the result is highly dependent on the final subjective decisions of the researcher. Therefore, improvements to our approach can be made through various means. When working in a group of experts assessing the categories, subjective bias could be reduced. The underlying taxonomy could also be revised and evaluated to ensure it fits the given context and application. In 2014 a framework was introduced that enabled a form of interactive topic modeling, which was structured in an iterative process, improving the topic model and reworking it based on user feedback throughout multiple cycles (Hu et al., 2014). This idea could be applied to our approach in further research. In the user study, multiple individuals noted that they missed certain categories. Feedback like this could be incorporated in future iterations to expand the completeness of the category system.

## 6.2 User Study

The results of the user study show a general consensus and direction of opinions. All users preferred the hierarchical organization of categories and were very vocal about the lack of it in systems 1 and 2. Furthermore, they preferred 70 to 90 topics and generally were happy with more specific topics and lower levels of abstraction. The disliking or critique regarding systems 1 and 2, for the most part, was directed at organizational aspects and stemmed from the users having difficulty navigating through the categories. Almost every participant said that the lack of grouping and thematic sorting led to the systems feeling "random" and "unorganized". Critique regarding the actual topic words was less frequent in comparison. For the most part, it was directed at terms that were "too general" or terms that participants had difficulty understanding and had trouble imagining which sorts of radio contributions would fall under the category.

Further comments on the content arrangement were the critique of mixing formats and topics or mixing topics of different granularity/abstraction. The users generally preferred when topics on the same logical level were coherent, of similar abstraction and type (as in topic or format). Based on this research, we can derive the following design principles:

1. The system should cover a broad range of categories with a total of 70 to 90
2. The system should contain two hierarchical levels covering two levels of abstraction
3. The hierarchical levels should be coherent and describe similar levels of abstraction as well as type

As our study was conducted from a user-focused perspective, those findings need to be considered with certain trade-offs in mind. The preference for lower levels of abstraction for a more precise specification of interest, for example, comes with the risk of having to continually add many categories to assure the completeness and an accurate representation of what contributions are available in the dataset. With a category like "Sports" for example, it covers all potential contributions regarding sports. When dividing it into more specific subcategories like "Soccer", "Tennis" etc.

more categories would have to be added to enable the user to select for all potential radio contributions. If a new document about chess or surfing is added, for example, and those topics have not been part of the dataset prior, they would not be covered by a finely separated categorization system anymore, whereas they would in the case of a coarse one. A potential workaround for this issue is adding an "other" category that includes outlier topics. Another caveat to be considered when trying to satisfy the user's preference for more options (more categories) and higher degrees of the specification is the danger of filter bubbles. They are encapsulated states of information created by algorithms or over-specification where a subject is isolated from diverse viewpoints and contents (Pariser, 2011). Not only do those echo chambers pose a threat to the quality of information and civic discourse, but they may also not be in the interest of the user himself. When presented with a narrow stream of topics, the user might get bored due to the lack of variation and potentially surprising content. To test this theory in further research, a long-term study should be conducted where the categorization systems are fully implemented into an application to test whether users still prefer the fine differentiation of topics long-term after listening to a highly personalized radio stream.

The fact that all participants liked the third hybrid system the most and preferred the automated over the manual system shows that automated topic modeling approaches can match and even exceed traditionally manual methods for topic modeling in quality and user satisfaction. The primary benefit stemming from this finding is the potential resource and time savings that come with the use of automation. Traditionally an expert would have to work through entire databases depending on the availability of metadata, which is a highly time-consuming task considering that most media publishers have years worth of historical data gathered. BERTopic, on the other hand, can generate an overview of the contents in a couple of minutes. Another benefit of automated topic modeling is the inherent classification of documents regarding topics. If the metadata is lacking, automated topic models can classify documents with their corresponding topics instantly, whereas this would require vast amounts of labour and resources if done manually. In the course of this study, we did not research this classification aspect or measure classification accuracy. Therefore, further research in this area needs to be conducted. Furthermore, it needs to be considered that with

BERTopic, there exist unclassified documents (17% of documents in our case), which are therefore not represented in the topic model.

The most prominent application we foresee is the hybrid approach, where the topic model acts as an assisting tool, as it combines benefits of both worlds. At the current stage, full automation produces still too unstructured results and needs post-processing to be presentable. Nonetheless, automation can help with the most labour-intensive portion of the process, identifying topics in the dataset.

This hybrid approach can exist in various proportions of manual to automated tasks. Depending on the use case, an almost fully automated version might be most suitable, where the human expert only makes minor corrections, cleans up, and adjusts for the potentially unclassified documents. Such a topic modeling with high degrees of automation might be most fitting for contexts where the topic collection requires frequent revision and updates and does not run the risk of creating a filter bubble as the automated methods tend to produce highly specific topics. An example of this might be the display of word clouds on a web page or blog of topics available. On the other end of the spectrum, a mostly manual approach is possible, during which the topic model is only used as a research tool for the expert to get an overview of topics in the data corpus. The categorization system is then constructed manually from scratch based on their knowledge and expertise. In applications where high degrees of structure and frequent user interactions are necessary, this approach might be the most suitable, for example, taxonomies or category lists for preference selection.

We faced certain limitations in our study which need to be considered and potentially improved upon in further research in this area. First of all, as noted prior, we were limited by time and resource variables in the process of constructing the optimal automated topic modeling approach. Furthermore, we were only able to compare three approaches (LDA, Top2Vec, BERTopic) even though numerous new topic modeling approaches regularly arise in the field of NLP. In the context of the user study, we were only able to work with a limited group size of ten people. Ideally, the study should be repeated with a larger sample size of participants to provide more representative results. Utilizing the think-aloud evaluation method comes with the disadvantage that we were not able to ask specific questions but had to keep the interview general and not focused on a certain aspect. The time intensive nature of the interviewing method also does not make it feasible for use in a large-scale study. Thus,

a different evaluation method like a survey or questionnaire with more pointed questions might be a better approach in future research. Additionally, a longer-term study would be of the highest interest to analyze whether the user perception and preferences regarding a topic modeling system change throughout long-term use. Due to the nature of our evaluation method, we were only able to capture the first impressions of the users. We are thus not able to generalize the findings in a long-term context confidently. Repeating the study with a large user base for a longer time with a deployed app that is already in use would be ideal. A final potential limiting factor might have been user interface design. We tried to work with a minimalistic and inconspicuous UI not to interfere with users' opinions. Nonetheless, research should be conducted on whether a particular UI design might significantly change users' preferences.

# 7 Conclusion

The rapidly growing amount of data and the increasing need for metadata as a basis for emerging technologies such as recommendation systems pose a major challenge for companies. We contributed by showing that automated solutions for topic modeling can be time and resource-efficient alternatives for manual methods. Furthermore, we presented the potential in a hybrid solution leveraging the strength of the automated and manual methods. We derived design principles for categorization systems through the user study and answered the research question based on direct user feedback.

As our study was focused mainly on user perception, a more holistic extension of our research problem addressing the limitations mentioned above is necessary to gain deeper insights into the potential and limit of automation in the domain of topic modeling. We encourage researchers to continue researching this issue as its importance continues to grow constantly.

## References

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv:2008.09470 [cs, stat]*. http://arxiv.org/abs/2008.09470

Angelov, D. (2020). *Ddangelov/Top2Vec* [Python]. https://github.com/ddangelov/Top2Vec (Original work published 2020)

Angular Components. (2021). *Angular Material*. Angular Material. https://material.angular.io/

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *Professional Communication, IEEE Transactions on*, *43*, 261–278. https://doi.org/10.1109/47.867942

Bruun, A., & Stage, J. (2015). New approaches to usability evaluation in software development: Barefoot and crowdsourcing. *Journal of Systems and Software*, *105*. https://doi.org/10.1016/j.jss.2015.03.043

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *arXiv:1707.01780 [cs]*. http://arxiv.org/abs/1707.01780

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336. https://doi.org/10.1145/290941.291025

Carrion, B., Onorati, T., Díaz, P., & Triga, V. (2019). A taxonomy generation tool for semantic visual analysis of large corpus of documents. *Multimedia Tools and Applications*, *78*(23), 32919–32937. https://doi.org/10.1007/s11042-019-07880-y

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, *22*. https://papers.nips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

*CUDA Toolkit*. (2013, Juli 2). NVIDIA Developer. https://developer.nvidia.com/cuda-toolkit

Debortoli, S., Müller, O., Junglas, I., & Brocke, J. vom. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, *39*(1). https://doi.org/10.17705/1CAIS.03907

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. http://arxiv.org/abs/1810.04805

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data, Rev. ed* (S. liii, 443). The MIT Press.

Grootendorst, M. (2020). *MaartenGr/BERTopic* [Python]. https://github.com/MaartenGr/BERTopic (Original work published 2020)

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165–181. https://doi.org/10.1080/01449290701773842

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*(3), 423–469. https://doi.org/10.1007/s10994-013-5413-0

Kotlerman, L., Avital, Z., Dagan, I., Lotan, A., & Weintraub, O. (2011). A Support Tool for Deriving Domain Taxonomies from Wikipedia. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 503–508. https://www.aclweb.org/anthology/R11-1069

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*. http://arxiv.org/abs/1405.4053

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(2), Article 2. https://doi.org/10.17169/fqs-1.2.1089

Mayring, P. (2014). *Qualitative content analysis: Theoretical foundation, basic procedures and software solution*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Musto, C. (2010). Enhanced vector space models for content-based recommender systems. *Proceedings of the fourth ACM conference on Recommender systems*, 361–364. https://doi.org/10.1145/1864708.1864791

*Natural Language Toolkit—NLTK 3.5 documentation*. (2021). https://www.nltk.org/

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The.

*Pretrained Models—Sentence-Transformers documentation*. (2021). https://www.sbert.net/docs/pretrained_models.html

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408.

https://doi.org/10.1145/2684822.2685324

*The Algorithm—BERTopic*. (2021). https://maartengr.github.io/BERTopic/tutorial/algorithm/algorithm.html

Urquhart, C. (2012). *Grounded Theory for Qualitative Research: A Practical Guide*. SAGE.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. https://doi.org/10.1145/1553374.1553515

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. 11.

# Appendix

Appendix 1: Categories Base vs. Data vs. Hybrid

| Basis | Data | Hybrid |
|---|---|---|
| Foreign Countries | Afghanistan | **Current** |
| Education | Africa | Brexit |
| Books | Astronomy | Coronavirus |
| Germany | Foreign Policy | Refugee Crisis |
| Docu & Feature | Automobile | Climate Crisis |
| Nutrition | Car Industry | **Foreign Countries** |
| Europe | Train | Africa |
| Movie | Balkans | America |
| Humanities | Banks | Asia |
| History | Basketball | Australia |
| Society | Education | China |
| Health | Biology | Europe |
| Radio Play | Black Lives Matter | Middle East |
| Interview | Brexit | USA |
| Children | Federal Policy | **Education** |
| Comments | Champions League | **Formats** |
| Culture | China | Biography |
| Market & Consumer | Comics | Docu & Feature |
| Media | Coronavirus | Radio Play |
| People | Democracy | Interview |
| Music | Digitalization | Comment |
| News | Doping | News |
| Internet | European Policy | Podcast |
| Politics | FIFA | Report |
| Press Review | Movie | Talk |
| Law | Finance | **Leisure** |
| Travel | Flora Fauna | Gaming |
| Religion | Refugee Crisis | Travel |
| Report | Frankfurt | **History** |

| | | |
|---|---|---|
| Sexuality | France | Holocaust |
| Sports | Soccer | **Society** |
| Talk | Gaming | Digitalization |
| Theatre | Money | Humanities |
| Environment And Transport | History | Philosophy |
| Entertainment | East & West Germany | Sexuality |
| Economy | Great Britain | **Culture** |
| Science | Holocaust | Visual Arts |
| | Home Office | Movie |
| | Hong Kong Protests | Literature |
| | Domestic Policy | Music |
| | Internet | Theatre |
| | Islam | **Media** |
| | Israel | Internet |
| | Japan | Journalism |
| | Journalism | **Nature** |
| | Judaism | Biology |
| | Church | **Politics** |
| | Classical Music | Foreign Policy |
| | Climate Protection | European Policy |
| | Art | Domestic Policy |
| | Artificial Intelligence | **Religion** |
| | Countries Policy | Christianity |
| | Literature | Islam |
| | Lockdown | Judaism |
| | Medicine | **Sports** |
| | Middle East | Basketball |
| | Music | Soccer |
| | Music & Festival | Handball |
| | National Socialism | Cycling |
| | Austria | Minority Sports |
| | Philosophy | Tennis |

| | | |
|---|---|---|
| | Podcast | Winter Sports |
| | Politics | **Technology** |
| | Press | **Transportation** |
| | Psychology | **Economy** |
| | Radio | Finances |
| | Cycling | **Science And Research** |
| | Minority Sports | Astronomy |
| | Space Travel | |
| | Religion | |
| | Russia | |
| | Sexuality | |
| | Skiing | |
| | Smartphones | |
| | Social Media | |
| | Spain | |
| | Spd | |
| | Sports | |
| | South America | |
| | Talk | |
| | Technology | |
| | Tennis | |
| | Terrorism | |
| | Theatre | |
| | Tourism | |
| | Trump | |
| | Turkey | |
| | Ukraine | |
| | USA | |
| | Venezuela | |
| | Belarus | |
| | Economy | |
| | Youtube | |