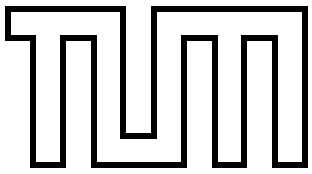


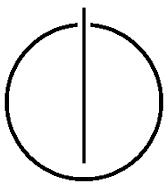
DEPARTMENT OF INFORMATICS

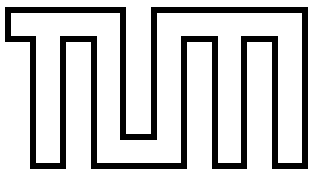
TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

# Topic Model Visualization for Opinion Mining

Maria Potzner






DEPARTMENT OF INFORMATICS

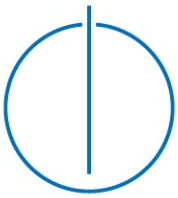
TECHNICAL UNIVERSITY OF MUNICH


Bachelor's Thesis in Information Systems

# Topic Model Visualization for Opinion Mining

# Topic Model Visualisierung für Opinion Mining


| | |
|---|---|
| Author: | Maria Potzner |
| Supervisor: | PD Dr. Georg Groh |
| Advisor: | PD Dr. Georg Groh |
| Submission date: | 15. Dezember 2018 |

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 15. Dezember 2018 Maria Potzner

# Abstract

Topic modeling is a popular approach to study large document collections. It returns topics as a set of coherent words, that are usually manually labeled, regarding the concept that the top 10 words of a topic are describing. In this work is examined how the labeling of topics can be automated. Further, the internal consistency is studied when the number of topics increases. Both approaches shall help domain experts using topic modeling for their work and understand which effect a higher or low number of topics has on the quality of the generated topics. To evaluate the results multiple datasets of online discussions regarding organic products are analyzed.

# Zusammenfassung

Topic modeling ist eine beliebte Methode große Textsammlung zu analysieren. Üblicherweise werden die identifizierten Themen als Liste an semantisch zusammenhängenden Wörtern dagestellt. Dieser Wörterliste wird dann manuell ein Betreff hinzugefügt. In dieser Arbeit werden unterschiedliche Ansätze verglichen, um diesen Prozess zu automatisieren. Des Weiteren wird die interne Konsistenz der Modelle bei der Veränderung der Topicanzahl analysiert. Beide Analysen sollen es Domänenexperten in Zukunft leichter machen Topic Modeling für ihre Arbeit einzusetzen. Außerdem, wird den Fachexperten dadurch ein besseres Verständnis für den Effekt der festzulegenden Topicanzahl vermittelt. Um die Ergebnisse zu überprüfen, werden die Analysen auf mehreren Datensätzen zu Onlinediskussionen über Bio-Produkten durchgeführt.

# Acknowledgement

As this thesis borders between computer science and qualitative research on consumer behaviour I would like PD Dr. Georg Groh of the Research Group for Social Computing for his input, support, good ideas and continuous feedback during the project and Hannah Danner from the Chair of Marketing and Consumer behavior. Without their collaboration this project and thesis would not be possible.

Furthermore, I would like to thank Gerhard Hagerer for his support and good ideas during the project and continuous feedback and Jan Hauffa, for his input regarding Topic Modeling.

As well i want to thank my parents Vaclava and Georg for their support during my Bachelor studies. Further, I would like to thank Christian Widmer for his support while working on this project and for proof-reading the thesis.

# Contents

# List of acronyms

# 1 Introduction

For researchers studying consumer opinions and trends user generated content is becoming an increasingly important input. Based on the discussions underneath online editorial sources or on discussion boards social scientists can perform opinion analysis on a scale that is not possible with classical approaches. The classical way of surveying consumers about their opinion on products or certain topics relies on voluntary questioning e.g. at supermarkets. These survey approaches have some drawbacks. First, it has to be made sure that the people questioned are representative for the larger population being studied. Second, in their response the participants might introduce a bias, since they know that they are being surveyed. Analyzing online user-generated data can help mitigate some of these drawbacks. Since the users do not know that their comments are used for opinion analyses there is less risk for bias. Further, any person has the ability to post online, which ensures that the comment sections and discussion boards contain a wide variety of opinions.

Even if user-generated content is used for qualitative studies, it is usually analyzed manually. With the huge growth of online text data, it is becoming of vital importance for social scientists to have reliable methods for fast automated analysis of such data. Among other things, researchers are interested in methods able to track topics, opinions, and sentiments in user-generated content (Nikolenko et al., 2017). Providing such a framework is the main objective of the *SocialROM* project.

From a Natural language processing (NLP) perspective finding topics in large document collections is known as topic modeling. Topic models take the documents as an input and outputs topics and for every document a distribution specifying how it is composed of these topics. In topic modeling, a topic is a probability distribution over all words in the documents. By ranking the words according to their probability every topic consists of different words. However, the words of one topic refer to the same concept or theme. For example, a topic with the most probable words *fish*, *salmon*, *wild* and *seafood* refers to the *fishing industry*. Each topic is a recurring theme that is discussed in the collection and is based on the co-occurrence of related words.

This project is done in cooperation with the chair of marketing research and consumer affairs at the Technical University of Munich. A part of their chair work

on qualitative methods for social media analysis and opinion elicitation regarding sustainable consumption and products. Therefore, our analysis of user-generated data focuses on discussions regarding organic products. This subject is fitting since organic vs. conventional food is a widely discussed online with diverse opinions and sentiments.

In *Generation 1* (Widmer, 2018) firstly all data, which were relevant for our domain regarding organic food and products were scraped. Further information can be found in chapter 3. Then topics were generated, with the focus on finding the best topics and showing them to our domain experts. To generate the best topics different parameters for Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were tried out and a method was developed to find the optimal number of topics per dataset. After creating the topics, a part of them was handed over to the domain experts and was labeled by them to evaluate, which datasets are meaningful. Based on the labels, the topic overlap of every dataset was considered, and it was discovered, that the discussed issues from editorial articles are more similar to the topics identified in forum threads than to the topics of editorial comments and the topics from blogs are most similar to blog comments. Furthermore, the topic labels were compared with the results of a qualitative survey, where people were asked why the buy organic products. The given reasons were also reflected in the topics derived from online discussions. Analogously, to Griffiths and Steyvers, 2004 the development of topics over time was considered, to identify trends.

This thesis builds up on the topics from *Generation 1*, theses were used to apply Automatic Topic Labeling (ATL) on them. The output of topic models are topics, which are represented with the top 10 words, sorted according the highest probability. It is desired, that the topics belong to a concept. For example from the top 5 words *costs, price, food, product* and *supermarket* it becomes apparent, that it is dealing with *food prices*. Therefore, the label food prices is assigned to the topic. The label assignment has so far been only done by domain experts, which is very time consuming for them. Therefore, different procedures for the automatic allocation of labels to topics were tried out and compared, in order to relieve the domain experts or to support them in their work. This is also necessary if topic modeling is generated on a growing live corpus and new topics can be constantly added, e.g. online discussions, and the actual themes shall be shown.

After ATL another main goal of this work is to prove the internal consistency, which means, to analyze how the topics itself and the distribution of topics on documents change when increasing the number of topics. Concretely, it shall be analyzed with different key figures whether the topics are getting specific or general and what specific and general in this context means. Furthermore, it shall be examined if the topics split up and whether this can be proven according to the top 10 words of

the topic. All theses analyses shall provide domain experts the overview how topics change when increasing or decreasing the number of topics, so they can assess, which topic model is the most appropriate one for their expected results.

## 1.1 Thesis structure

First the methods, which were used to identify the topics are introduced in Chapter 2. This includes the approaches to represent the content of documents numerically and the algorithms for topic modeling with Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) . In chapter 3 we introduce the dataset and show how the data were gathered and preprocessed. In the first half of Chapter 4, in Section 4.1, the possible approaches for Automatic Topic Labeling are described and the results of applying those on our dataset are discussed. Accordingly, in the second half of Chapter 4, in Section 4.2, different key figures to measure the internal consistency are first introduced, applied and then discussed on our dataset. Chapter 5 completes the thesis by providing an outlook for possible future work and summarizes the thesis with the conclusion.

# 2 Methodology

In this chapter the basic principles for the following chapters will be explained. The Section 2.1 describes how documents can be numerically represented. Section 2.2 then will introduce the topic modeling methods LDA and NMF which are used in this thesis.

## 2.1 Document representation

### 2.1.1 Bag of Words

The Bag of Words Bag of Words (BoW) model serves as a numerical representation of a document, which is used as input for further NLP tasks. It represents the document simply by the counts for each word. The grammar and the ordering of the words are ignored, so some information is lost. The document *John likes organic but Mary doesn't* and the document Mary likes organic but John doesn't have the same BoW representation although these differ in context. Nevertheless, similar BoW imply similar document content (Manning et al., 2008).

### 2.1.2 Tf-Idf Weighting

Only considering the absolute term frequency ($tf_{t,d}$) of words is not the best measure to make differentiations between documents, because not all terms are equally important. The term *organic* appears in 224 of 239 articles in the New York Times, obviously this term can not be considered as a stop word, however it is not suitable to differentiate the articles. Therefore the effect of the frequent words is reduced by the *inverse document frequency*:

$$idf_{d,t} = log \frac{N_d}{df_{d,t}} \tag{2.1}$$

$N_d$ is the number of all documents in a corpus, while $df_{d,t}$ is the number of documents that contain the single term.

Based on the term frequency $tf_{t,d}$ and the inverse document frequency $idf_{d,t}$ we introduce the *term frequency - inverse document frequency (tf-idf)*:

$$tf - idf_{d,t} = tf_{t,d} * idf_{t,d} \tag{2.2}$$

The tf-idf weighting has the highest score when the term occurs frequently within a small amount of documents. The score is lower when the term occurs rarely or too often in many documents (Jurafsky and Martin, 2009).

### 2.1.3 Vector space model

The representation of documents in the same vector space is known as the vector space model. This was originally introduced for Information Retrival (IR) operations like scoring documents on a query, document classification or clustering Salton et al., 1975.
The vector space model forms with the documents $D_i$ and all unique terms $T_j$ the document term matrix *C*. Each row of *C* corresponds every single document of the corpus and each column the single unique terms. In $C_{ij}$ the weightings either as term frequency or tf-idf for each term over all documents is stored.
In Table 2.1 the term frequency and in Table 2.2 tf-idf is calculated from three sample documents: *Doc 1: Organic is healthier then conventional food*, *Doc 2: I buy organic* and *Doc 3: Organic is wasted money*. In this thesis both topic modeling algorithms take the document term matrix as input, but with different weightings. For LDA the term frequency and for NMF the tf-idf weighting is used.

| | organic | is | healthier | then | conventional | food | i | buy | wasted | money |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| Doc2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| Doc3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Tab. 2.1.:** Document term matrix with term-frequency weighting as used by LDA.

| | organic | is | healthier | then | conventional | food | i | buy | wasted | money |
|---|---|---|---|---|---|---|---|---|---|---|
| Doc1 | 0 | 0.45 | 0.45 | 0.45 | 0 | 0.34 | 0 | 0.27 | 0.45 | 0 |
| Doc2 | 0.65 | 0 | 0 | 0 | 0.65 | 0 | 0 | 0.39 | 0 | 0 |
| Doc3 | 0 | 0 | 0 | 0 | 0 | 0 .44 | 0.58 | 0.34 | 0 | 0.58 |

**Tab. 2.2.:** Document term matrix with tf-idf weighting as used by NMF.

## 2.2 Topic Modeling

Every day large amounts of information are collected and become available. The vast quantities of data make it difficult to access those information we are looking for. Therefore, we need methods that help us to organize, summarize and understand large collections of data. Topic Modeling refers to a set of methods that help us to process large document collections efficiently. A topic model takes a set of documents as the input and outputs topics, a set of recurring themes that are discussed in the collection, and the degree to which each document expresses these topics (Blei, 2003). The topics are the hidden thematic structure of the document collection. They are found by recurring patterns of co-occurring words.

Topic Models are based on the assumption that a document is composed of multiple topics. Any document of the text collection is a combination of all topics with different weights. Therefore, all documents of one corpus are composed of the same topics by varying their weights. The documents are modeled as a probability distribution over the topics.

Just as documents are distributions over all topics, topics are distributions over all words in a document collection. Every word of the collection is present in every topic, however, with varying probabilities. Sorting the terms of a topic by their probability reveals a semantically meaningful interpretation.

### 2.2.1 Latent Dirichlet Allocation

Currently, the most used method for topic modeling is Latent Dirichlet Allocation, which was introduced by Blei, 2003. LDA is a generative model, that describes how the documents are generated from existing topics. By applying inference it is possible to reverse the process and use LDA to derive the topics from the document collection. Further, LDA is also a probabilistic model which means that the resulting topics can be seen as a probability distribution.

The research on topic models was started by Deerwester et al., 1990 with their introduction of Latent Semantic Analysis (LSA). LSA was used to solve the issues of polysemy and synonymy when performing queries for information retrieval. With probabilistic LSA (pLSA) Hofmann, 2001 provided a probabilistic formulation of LSA. LDA is an extension of pLSA to be able to model unseen documents. In pLSA the topic proportions for every document need to be known, which makes it unable to model documents outside the training set. In LDA the topic proportions for every

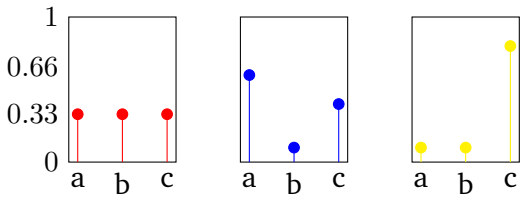



**Fig. 2.1.:** Discrete distributions drawn from a 3-dimensional Dirichlet distribution. Adapted from Widmer, 2018

document are derived from a Dirichlet distribution, thus enabling it to generalize to documents outside the training collection.

After the successful introduction to analyze large document collections, LDA was also applied in other domains such as computer vision to classify (Fei-Fei and Perona, 2005) and build hierarchies of images (Li et al., 2010). In fact, the method was independently discovered by Pritchard et al., 2000 in their field of evolutionary biology to study population genetics. For the interested reader, Jelodar et al., 2017 presents a literature study on extensions of LDA and applications to various datasets.

As described above LDA assumes a generative process how the documents of a corpus arose from existing topics. We first describe the generative process before explaining how it can be reversed to derive topics from existing documents.

LDA draws the per-document topic distribution and the topics themselves from a Dirichlet distribution. Therefore, before explaining LDA the Dirichlet distribution is introduced.

The Dirichlet distribution can be seen as a distribution over distributions. When sampling from a k-dimensional Dirichlet distribution one receives a discrete distribution over k elements. This is visualized in Figure 2.1. Each distribution represents a draw from a 3-dimensional Dirichlet distribution.

Further, the Dirichlet distribution can be parametrized by $\alpha$. The parameter describes how the probability mass of the Dirichlet distribution is distributed over the k-elements. When drawing from a Dirichlet with a high alpha parameter the probabilities for each element of the drawn discrete distribution are roughly the same. When drawing from a Dirichlet with a small alpha, the probability mass of the resulting discrete distribution is divided among a few highly probable elements. Visually, a high $\alpha$ value leads to a distribution as shown on the left of Figure 2.1. A low value would lead to distributions as shown on the right.

Figure 2.2 represents LDA as a graphical model. The grey nodes indicate observed variables, in this case the words $w$ of the documents. All white nodes are hidden variables that have to be derived. In topic modeling the hidden variables are the topic assignment $z_{w,d}$ of each word position $n$ in document $d$, the topic distribution $\theta_d$ for document $d$ and the word distribution $\phi_z$ for topic $z$. $\alpha$ and $\beta$ are hyper-parameters for the Dirichlet distribution on the per-document topic distribution and per-topic word distribution respectively. With the introduced notation the generative process underlying LDA can be described as follows:

- From a Dirichlet distribution parametrized by $\alpha$ draw a multinomial distribution $\theta_d$ representing the topic proportions of document $d$.
- For each word position $n$ in document $d$ choose a topic of the multinomial per-document topic distribution $\theta_d$. The chosen topic is the topic assignment $z_{n,d}$.
- From a Dirichlet distribution parametrized by $\beta$ draw a multinomial distribution $\phi_z$ representing the word proportions for topic $z$.
- The word $w$ at position $n$ in document $d$ is then drawn from the topic $z$.

By applying this procedure for all documents and words we can generate the documents from existing topics. To derive the topics from existing documents we need to estimate the document-topic proportions $\theta$, the topics $\phi$ and the assignment of words to topics $z$ given the Dirichlet priors $\alpha$ and $\beta$ and the word of the documents $w$. This can be formulated as:

$$P(\theta, \phi, z|w, \alpha, \beta) = \frac{P(\theta, \phi, z, w|\alpha, \beta)}{P(w|\alpha, \beta)} \tag{2.3}$$

This fraction, however, is intractable to compute. Therefore, several approaches exist, such as Gibbs Sampling (Griffiths and Steyvers, 2002), Variational Inference (Blei, 2003) or Expectation Propagation (Minka and Lafferty, 2002), to approximate the topic-term and document-topic distributions.

### 2.2.2 Non negative Matrix Factorization

Apart from the probabilistic methods as described above, linear methods, such as Non-negative Matrix Factorization have proven useful for topic modeling. NMF was introduced by Lee and Seung, 1999 as a method for dimensionality reduction. They show that their method can lead to a lower dimensional parts-based representation

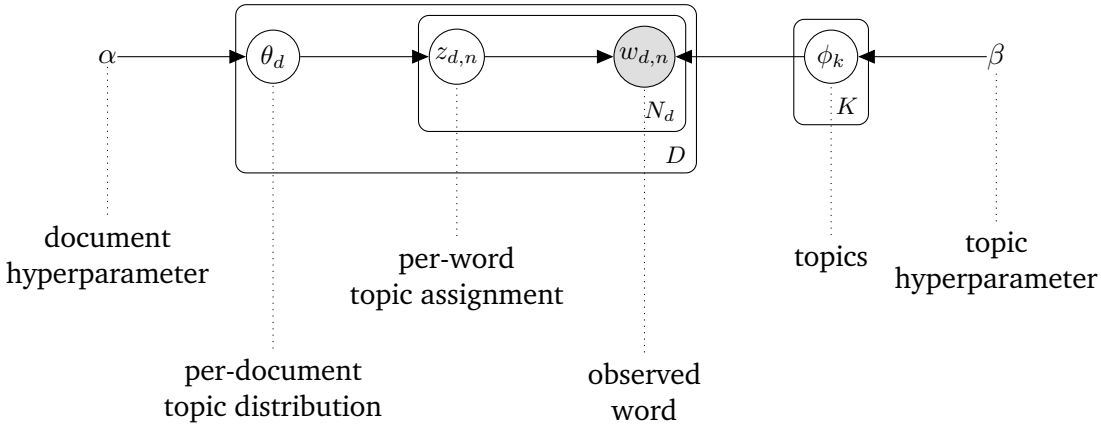



**Fig. 2.2.:** Graphical model of LSA. Adapted from Widmer, 2018

with naturally interpretable components. When applied on a set of facial images, the parts resemble different portions of a human face. The original facial images can be reconstructed by combining the parts. This approach can also be applied on a set of textual documents. In this case the parts-based representation are the semantic topics of the documents and the documents can be reconstructed by combining the topics.

NMF requires that the original data is non-negative. This is the case for the previously presented applications in computer vision and text mining. The pixel values of images or term counts of documents are always positive. It is also typical to data in other fields that they are non-negative. Therefore, NMF was also successfully applied in spectral analysis (Pauca et al., 2006) and bio-informatics(Brunet et al., 2004).

Formally, NMF can be described as follows. Given a non-negative matrix $C \in \mathbb{R}_{\geq 0}^{n \times m}$ and a rank desired $k < min(m, n)$ find two non-negative matrices $W \in \mathbb{R}_{\geq 0}^{n \times k}$ and $H \in \mathbb{R}_{\geq 0}^{k \times m}$ so that the reconstruction error is minimized:

$$\min_{W,H} ||C - WH||_F \tag{2.4}$$

In this case, the reconstruction error is measured by the Frobenius norm, which is an extension of the Euclidean form on matrices. The output and input of NMF is similar to LDA. Matrix $C$ is the original document term matrix whereas the matrices $W$ and $H$ represent the topic term respectively the document topic matrix. For every document $c$ there is a corresponding column $h$ in the document topic matrix $H$. The column contains the weights for each topic (columns of $W$) for this specific document. Thus, the document $c$ is modeled as a linear combination of the columns of $W$:

$$c \approx Wh \tag{2.5}$$

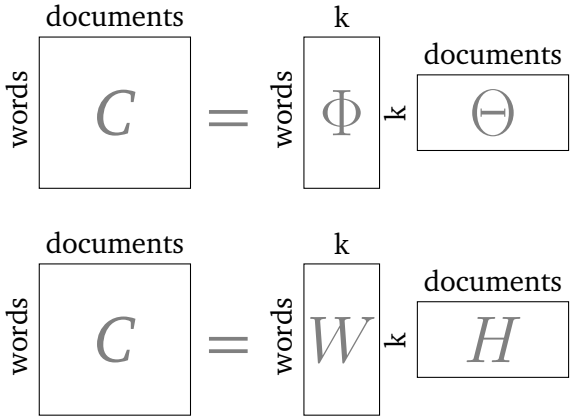



**Fig. 2.3.:** Showing the similarity of LDA, and NMF from the perspective of matrix decomposition. Adapted from Steyvers and Griffiths, 2007b.

. The similarity of LDA and NMF in terms of matrix factorization is evident from Figure 2.3. However, there is an important difference in the interpretation of the values in the generated matrices. With LDA the outputted document-topic and topic-term matrices are probability distributions. NMF, however, has no probabilistic interpretation and while returned values represent topic or term weights they do not necessarily sum up to 1. To circumvent this difference the output of NMF was normalized.

# 3 Dataset

In order to identify and analyze the consumers decisions in context of sustainable food we need a large dataset, which consists of different sources to capture the various opinions and discussion topics of the large population. The following chapter summarizes how the relevant datasets of editorial resources, personal blogs and discussion boards were selected and preprocessed in *Generation 1* and which changes were made. Afterwards it is described how the topics of the datasets were identified. Based on already existing and new generated topics together with the scraped datasets, the following chapters presents further analysis and additional insights.

## 3.1 Data collection

To gather a wide rage of opinions towards sustainable food and the variation of discussion topics over time, different datasets such as online editorial news sites, blogs and discussion boards were considered in the period from January 2007 until November 2017. These datasets are all public and without any charge available online. Additionally, the user generated data, such as comments under articles or in forums, can be posted by using a pseudonym and the users do not know their data will be studied. This reduces the potential of response bias, which is usually present when performing surveys or experiments.

Online outlets of supra-regional print press, national print press (IVW, 2018)[1] and the news sites (AGOF, 2018)[2] were selected according to the highest reach by the Domain experts. Blogs and forums were selected with the help of snowball technique, meaning Domain experts‘ colleagues identified further sustainable blogs or forums. This kind of data were selected for Germany, Austria, Swiss and the US.

After the selection, the chosen datasets were automatically scraped and examined for terms like *bio Lebensmittel*, *bio Landwirtschaft* for the German and terms like *organic, organic food, organic agriculture*, and *organic farming* for the English language using site‘s internal search engines or Google search, which offers the option to search for sites within a domain. Nevertheless, still non relevant data like recipes, product

[1] only an example German national print press
[2] only an example German news sites

presentations, and stock market information remained. These were kicked out by the binary Naive Bayes classifier, which was trained on 1000 random articles[3], that were labeled either as relevant or not by the Domain experts. The final collection stored in a JSON schema and the list of all sources and their percentage of relevant articles together with other descriptive statistics can be found in Appendix A.

## 3.2 Data processing

For applying further NLP tasks, the extracted dataset was transformed by using several pre-processing tasks: First, the texts were tokenized and lowercased. Then all common words including numbers and punctuations were removed and Emails and Url's were replaced by <EMAIL> and <URL> tags. Second, the remaining tokens were lemmatized, so that the inflections of words were replaced by their basic form. Third, the texts were examined for collocations, which are co-occuring words like *Stiftung Warentest* or *Whole Foods*, with a Gensim library [4]. For the lemmatization and tokenization the Spacy library [5] was used. Additionally, in this project Part-of-speech (POS)-Tagging was applied to the texts, which is a process marking up the words to a particular part of speech, to facilitate the ATL in chapter 4.1.

## 3.3 Final Datasets

Before reporting the datasets itself, the definition of text types will be described, which were introduced because of the different content and language style. All data referring to a main text of a side are called *editorial articles* and the comments under the editorial articles are called *editorial comments*. The term *Forum* includes the initial question and the comments under it. In this thesis the blogs, which were split in editorial and comments, were neglected, because the amount of data and context quality was to low.

We created two different final datasets where the frequent words, occurring over 90% in a document, and the infrequent words, occurring under 0,05%, were kicked out. The first dataset consists of editorial articles, editorial comments and forums. The final number of documents and amount of words is listed in Table 3.1. The second dataset consists of editorial articles and the summarized comments from the editorials and forums. This is shown in Table 3.2. Both datasets were built for the German and English language.

[3] contains the title, text and text of 100 comments
[4] https://radimrehurek.com/gensim/index.html
[5] https://spacy.io

| | | Editorials | | Forums |
|---|---|---|---|---|
| | | articles | comments | |
| German | # documents | 4730 | 1782 | 641 |
| | # words | 5239 | 15413 | 7361 |
| English | # documents | 2345 | 441 | 3274 |
| | # words | 6254 | 11948 | 5970 |

**Tab. 3.1.:** The number of documents and vocabulary sizes for Editorials and Forums of the German and English datasets.

| | | Editorial articles | Comments |
|---|---|---|---|
| German | # documents | 4730 | 2423 |
| | # words | 5239 | 22774 |
| English | # documents | 2345 | 3715 |
| | # words | 6254 | 17918 |

**Tab. 3.2.:** The number of documents and vocabulary sizes for Editorial articles and Comments of the German and English datasets.

## 3.4 Topic Generation

The complete dataset not only includes the texts but also topics, that were identified as part of *Generation 1*. These topics were generated separately by language and text type. Since we merged comments underneath editorial articles and forums, we generated new topics based on the same parameter and the same approach to select the number of topics. Generating qualitative topics depends on the hyper parameters $\alpha$ and $\beta$ for LDA and the topic number for LDA and NMF. The domain and the documents influence the optimal values for the hyper parameters. Therefore, in *Generation 1*, the $\alpha$ and $\beta$ were determined by analyzing the topic coherence and the perplexity of the topics. The asymmetric $\alpha$ and symmetric $\beta = 0.01$ were considered as the best values. These were used to generate the previous topics and the new ones for summarized comments. Obtaining the best topic number for each dataset multiple Topic Models were trained for a range of different number of topics with LDA and NMF. The following steps describe the process to estimate the optimal number of topics for a language, dataset and algorithm e.g. English Comments with NMF:

1. For every Topic Model with different topic numbers a plot was generated, see Figure 3.1. The x-axis shows the values for the most probable topic for every single document while the y-axis shows the counted documents where the topic occurs.

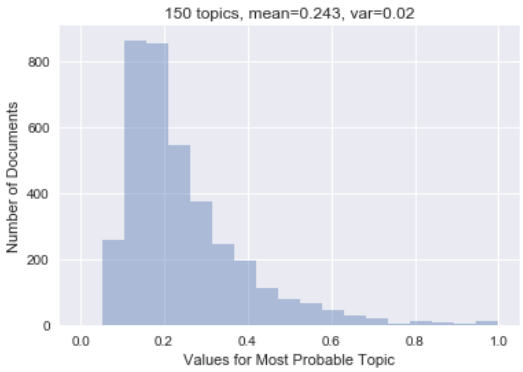



**Fig. 3.1.:** Count of the value of the most probable topic, summed over all topics.

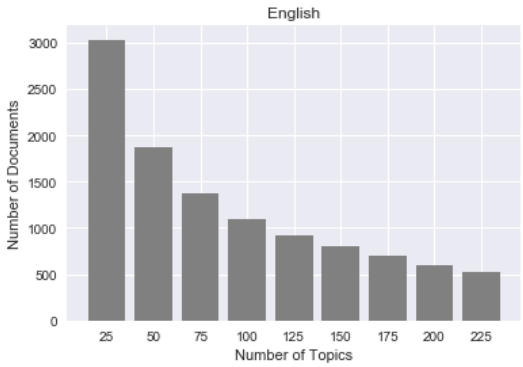



**Fig. 3.2.:** Number of documents the topics are expressed above the threshold

2. In each plot the mean of the x-axis values was calculated. Afterwards the means of all plots were averaged and used as a threshold in the next step.

3. The number of documents was summed up if the probability of the topics was greater then the threshold. The sum was calculated for every Topic Model and plotted in Figure 3.2.

4. The point where the curve flattens, was taken as the optimal topic number.

After finding the appropriate topic number, the Topic Models generated with NMF and LDA for the same dataset were inspected manually. The domain experts labeled the topics and the Topic Model with the higher number of labels was chosen. The final selection of the Topic Models including also Topic Models with summarized comments is shown in Table 3.3.

| | Editorials | | | | Forums |
|---|---|---|---|---|---|
| | Editorial articles | Comments | articles | comments | |
| German | **190** | *125* | **190** | *170* | *170* |
| English | **130** | *125* | **130** | *170* | **110** |

**Tab. 3.3.:** The optimal number of topics for Editorials and Forums. *Italic* denotes NMF and **bold** numbers denote LDA.

# 4 Experiments and Results

## 4.1 Automatic Topic Labeling

Topic Models are used to discover latent topics in a corpus to help to understand large collections of documents. These topics are multinomial distributions over all words in a corpus. Normally, the top terms of the distribution are taken to represent a topic, but these words are often not helpful to interpret the coherent meaning of the topic. Especially, if the person is not familiar with the source collection. For example, for the topic *price, $, cost, foods, store, product, brand, low, supermarket, good* a suitable label is *food prices*.

With the help of Automatic Topic Labeling (ATL) we want to reduce the cognitive overhead of interpreting these topics and, therefore, facilitate the interpretation of the topics. Of course, the topics can be labeled manually by domain experts, but this method is time consuming if there are hundreds of topics. Additionally, the topic labels can be biased towards the users opinion and the results are hard to replicate.

We are working with domain specific data dealing with organic food. To generate meaningful labels we can not make use of human turks because we need domain experts who are proficient in this area. Therefore, we submitted the topics to our domain experts to label them. But only 50 of the generated topics, ranked according to the importance in a corpus, for each dataset were handed in, in order to not burden them, since this process is very time-consuming. The datasets were labeled by three labelers who tried to find a suitable label, which captures the meaning of the topic and is easily understandable. After labeling, the three labels of a topic were compared and a final label was set. If at least two labelers had the same label, this was taken as the final one. If the given labels were not comparable, no label was set at all.

To relieve our domain experts in the following chapter two approaches for ATL are described. In Section 4.1.2 an intrinsic method was used, which is only working on texts and topics from our datasets to generate the labels according Mei et al., 2007. Section 4.1.3 describes an extrinsic approach by using a lexical database for the English language called *Wordnet* to label the topics.

### 4.1.1 Related work

*Lau et al., 2011* generated a label set,called primary candidate labels, out of article titles, which were found in Wikipedia or Google by searching the top N words from topics. Afterwards, these labels were chunkized and n-grams were generated. Theses secondary candidate labels were then filtered with the *related article conceptual overlap* (RACO), that removed all outlier labels, such as stop words. Then the primary and secondary candidate labels were ranked by features such as point-wise mutual information (PMI), used for measuring association, and the student's t test. The results were measured with the mean absolute error score for each label, which is an average of the absolute difference between an annotator's rating and the average rating of a label, summed across all labels. The score lay between 0.5 and 0.56 on a scale from 0 to infinity.

On topics from Twitter *Zhao et al., 2011* used a topic context sensitive Page Rank to find keywords by boosting the high relevant words to each topic. These keywords were taken to find keyword combinations (key phrases) that occur frequently in the text collection. The key phrases were ranked according to their relevance,i.e. whether they are related to the given topic and discriminative, and interestingness, the re-tweeting behavior in Twitter. To evaluate the keywords Cohen's Kappa was used to calculate the iterrater reliability between manually and automatically generated key phrases. The Cohen's Kappa coefficient was in the range from 0.45 to 0.80, showing good agreement.

*Allahyari and Kochut, 2015* created a topic model OntoLDA which incorporates an ontology into the topic model and provides ATL too. In comparison with LDA, OntoLDA has an additional latent variable, called concept, between topics and words. So each document is a multinational distribution over topics, each topic is a multinomial distribution over concepts and each concept is a multinomial distribution over words. Based on the semantics of the concepts and the ontological relationships among them the labels for the topics are generated in followin steps:

1. construction of the semantic graph from top concepts in the given topic

2. selection and analysis of the thematic graph (subgraph form the semantic graph)

3. topic graph extraction from the thematic graph concepts

4. computation of the semantic similarity between topic and the candidate labels of the topic label graph

The top N labels were compared with the labeling from *Mei et al., 2007* by calculating the precision after categorizing the labels into good and unrelated. The more labels were generated for a topic the more imprecise they got but the preciser *Mei et al., 2007* labels were.

*Hulpus et al., 2013* made use of the structured data from DBpedia, that contains structured information from Wikipedia. For each word of the topic the Word-sense disambiguation (WSD) chose the correct sense for the word from a set of different possible senses. Then a topic graph was obtained form DBpedia consisting of the closest neighbors and the links between the correct senses. Assuming the topic senses which are related, lie close to each other, different centrality measures were used and evaluated manually to identify the topic labels. The final labels then were compared to textual based approaches and the precision after categorizing the labels into good and unrelated was calculated.

Kou et al., 2015 captured the correlations between a topic and a label by calculating the cosine similarity between pairs of topic vectors and candidate label vectors. Continuous bag of words (CBOW), Skip-gram and Letter Trigram Vectors were used. The candidate labels were extracted from Wikipedia articles that contained at least two of the top N topic words. The resulting labels for the different vector spaces were compared to automatically generated gold standard labels, representing the most frequent chunks of suitable document titles for a topic. The final labels were ranked by human annotators,too, and were considered as a better solution than the first word of the top N topic words.

For topics and preprocessed Wikipedia titles *Bhatia et al., 2016* used word and title embeddings. To generate title embeddings doc2vec and word2vec were used to obtain fine-grained labels (doc2vec) or generic labels (word2vec). Given a topic, the relevance of each title embedding was measured based on the pairwise cosine similarity with each of the word embeddings for the top-10 topic terms. The sum of of the relevance of doc2vec and vec2doc served as ranking for the labels. The results were evaluated the same way as like in Lau et al., 2011.

*Magatti et al., 2009* used a given tree-structured hierarchy from the Google Directory to generate candidate labels for the topics. These were compared to the topic words by applying different similarity measures. The most suitable label was then selected by exploiting a set of labeling rules. This approach is applicable to any topic hierarchy summarized by a tree.

*Mei et al., 2007* generated labels based on the texts collection and their related topics by chunking and building n-grams. They approximated the distribution for the labels and compared these to the distribution of the topic by calculating the

Kullback Leibler (KL) divergence. To maximize the mutual information between the label and the topic distributions the calculated divergence has to be minimized. Three human assessors measured the results and found out that the final labels are effective and robust although applied on different genres of text collections.

### 4.1.2 Intrinsic Topic Labeling

The intrinsic topic labeling is based only on a text collection and therefrom extracted topics.It does not use any external ontologies or embeddings. Because *Mei et al., 2007* were the only ones who generated topic labels by using an intrinsic approach, we decided to apply their ATL on our data, using an implementation from Github[1]. The implementation was adapted to our data and instead of using their preprocessing ours was used.

In their paper *Mei et al., 2007* consider noun phrases and n-grams as candidate labels and use Part-of-speech (POS)-tags to extract the labels according to some grammar from the text collection. We apply the n-grams approach to select (NN - NN) or (JJ - NN) English and (NN -NN) or (ADJD - NN) German bi-grams as suitable labels for the topics.

The candidate labels were ranked by their semantic similarity to the topic distribution $\theta$. To measure the semantic relevance between a topic and a label $l$ a distribution of words $w$ for the label $p(w|l)$ was approximated by including a text collection $C$ and a distribution $p(w|l, C)$ was estimated, to substitute $p(w|l)$. Then the Kullback Leibler (KL) divergence $D(\theta||l)$ was applied to calculate the closeness between the label and the topic distribution $p(w|\theta)$. So the KL divergence served to capture how well the label fits to the topic. If the two distributions perfectly match each other and the divergence is zero we have found the best label. The relevance scoring function

[1] https://github.com/xiaohan2012/chowmein

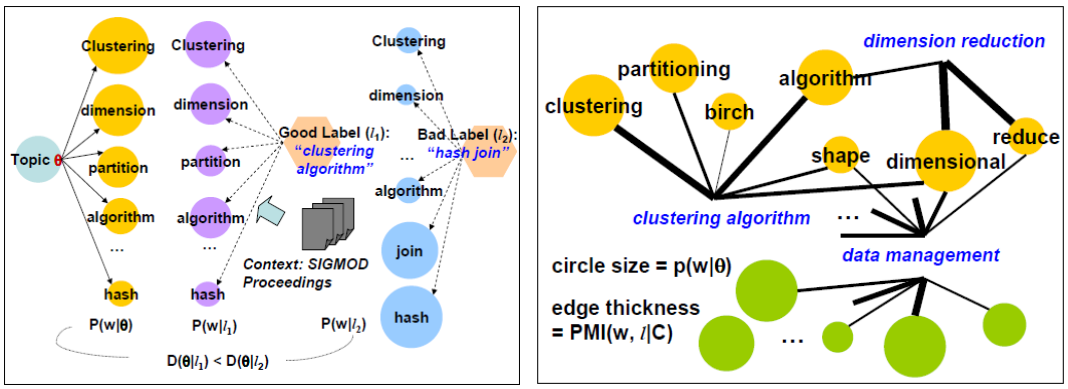



**Fig. 4.1.:** Relevance scoring function for ATL. Adapted from Mei et al., 2007

of $l$ to $\theta$ is defined as the negative KL divergence $-D(\theta||l)$ of $p(w|\theta)$ and $p(w|l)$ and can be rewritten as follows by including $C$:

$$\begin{aligned}
Score(l,\theta) &= -D(\theta||l) = -\sum_w p(w|\theta) log \frac{p(w|\theta)}{p(w|l)} \\
&= -\sum_w p(w|\theta) log \frac{p(w|C)}{p(w|l,C)} - \sum_w p(w|\theta) log \frac{p(w|\theta)}{p(w|l)} \\
&\quad -\sum_w p(w|\theta) log \frac{p(w|l,C)}{p(w|l)} \\
&= -\sum_w p(w|\theta) log \frac{p(w,l|C)}{p(w|C)p(l|C)} - D(\theta||C) \\
&\quad -\sum_w p(w|\theta) log \frac{p(w|l,C)}{p(w|l)} \\
&= -\sum_w p(w|\theta) PMI(w,l|C) - D(\theta||C) + Bias(l|C)
\end{aligned} \tag{4.1}$$

We can see that the relevance scoring function consists of three parts. The first part represents the expectation of PMI $E_\theta(PMI(w,l|C))$ between $l$ and the words in the topic model given the context $C$, the second part is represented by the KL divergence between $\theta$ and $C$ and the third part can be viewed as a bias using context $C$ to infer the semantic relevance $l$ and $\theta$. This bias can be neglected for our data because we have used the same text collection for producing the topics and the labels. The same applies to the second part, because the KL divergence has the same value for all candidate labels. Therefore, we rank the topic labels with

$$Score(l,\theta) = E_\theta(PMI(w,l|C)) \tag{4.2}$$

The relevance scoring function is also described visually in Figure 4.1. The circles represent the probability of terms. The larger the circle the higher is the probability. On the left one can see that the label with lower KL divergence is the best one. To approximate $p(w|l)$ in this example the *SIGMOD Proceedings* were used as the text

collection $C$, not in our implementation. Analogously, we used our datasets. On the right one can see a weighted graph, where each node is a term in the topic distribution $\theta$ and the edges between terms and the label are weighted by their PMI. The weight of the node indicates the importance of a term to the topic, while the weight of each edge indicates the semantical strength between label and term. The relevance scoring function ranks a node higher if the label has a strong semantic relation to the important topical words. Visually, this can be understood that the label is ranked higher if it connects to large circle by a thick edge.

So far only the labeling of a topic was considered, but a characteristic of a good label is the discrimination towards other topics in the topic model, too. It is not useful if many topics have the same labels, although it may be a good label for the topic individually, because we can not make differentiations between the topics. The label should have a high semantic relevance to a topic and low relevance to other topics. In order to take this property into account the $Score(l, \theta)$ in 4.2 was adjusted to:

$$Score'(l, \theta_i) = Score(l, theta_i) - \mu Score(l, \theta_{1,\dots,i-1,i+1,\dots}) \tag{4.3}$$

$\theta_{1,\dots,i-1,i+1,\dots}$ describes all topics except the $\theta_i$ and $\mu$ controls the discriminative power. In our implementation we set $\mu$ to 0.7.

### 4.1.3 Extrinsic Labeling

The majority of literature uses extrinsic topic labeling approaches, using external ontologies or data, because the achieved results are better than the ones from the intrinsic approach. Existing approaches working with e.g. Wikipedia, DBpedia and Google directory as used by *Lau et al., 2011*, *Hulpus et al., 2013*, *Bhatia et al., 2016* and *Magatti et al., 2009* are not applicable on our specific data. Therefore, we were looking for a method that can be applied on our domain-specific data.

We used the English online database *WordNet*[2], that contains 118.000 different word forms and 90.000 word senses. WordNet organizes the several types of words like nouns, verbs, adjectives and adverbs into sets of synonyms, called *synsets*. A *synonym* is a word that has the same meaning as another word. E.g *shut* is a synonym for *close*. These two words form together with possibly other words such as *fold* a synset. Additionally, a synset contains a short definition, called *gloss*, and an exemplary sentence for each term in a synset, which describes the usage of this term. Every distinct word sense of a given word is represented as a separate synset. So the number of different meanings for a word corresponds to the number of synsets. All synsets are linked to each other according to semantic relations such as synonymy,

[2] http://wordnetweb.princeton.edu/perl/webwn

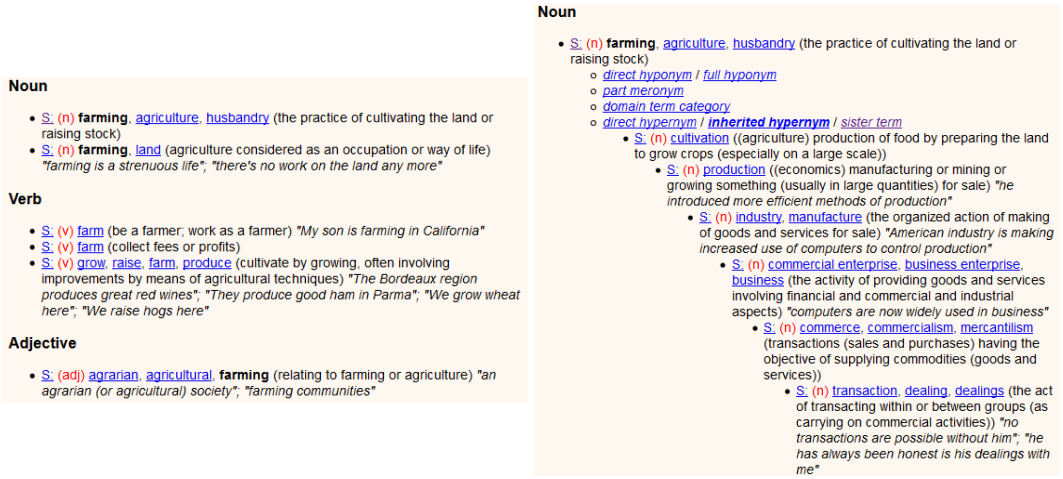



**Fig. 4.2.:** WordNet results for the word *farming*. Adapted from *WordNet*

antonymy, hyponymy, hepernymy, meronymy and troponymy. A definition of these semantic relationships can be found in Miller, 1995. In our implementation we used besides *synonymy* also *hypernymy*. If two words can be generalized by an other word, this word is called *hypernym*. E.g *animal* is a hypernym for *cat* and *dog*.

In Figure 4.2 one can see the resulting synsets when typing the word *farming* into WordNet. Synsets of nouns (*farming, agriculture, husbandry* and *farming, land*), verbs (two different meanings of *farm* and *grow, raise, farm, produce*) and adjectives (*agrarian, agricultural,farm*) were found, that can be seen on the left side. For each synset the inherited hypernym can be determined. An excerpt of inherited hypernyms(*cultivation,production, industry etc.*) for the synset *farming, agriculture,husbandry* is shown on the right. These are forming a hierarchical tree. The lower a hypernym in the tree the more general it is. In this figure the synset *production* is more general than synset *cultivation*. The most general or lower hypernym for all synsets in WordNet is *entity*.

To extract the information from WordNet we used the *NLTK corpus reader*.[3] In addition to WordNet also Polyglot [4] was used as kind of preprocessing for selecting similar words of a topic by using word embeddings.

### Preprocessing

For all following approaches in the next section we implemented a preprocessing step, that can be applied before running the different approaches for labeling a topic. It should improve the quality of the labels. Our topics consists of 10 words,

[3] http://www.nltk.org/howto/wordnet.html
[4] https://polyglot.readthedocs.io/en/latest/Embeddings.html

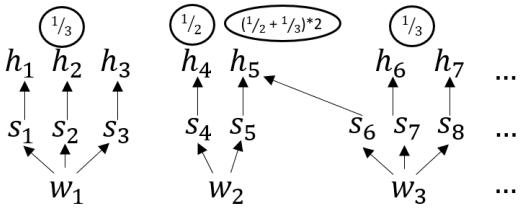



**Fig. 4.3.:** Scoring function for hypernyms

usually these words can not be summarized to one label, which fits to all of the topic words. Therefore, the distances between every combination of two topic words were calculated with Polyglot embeddings. The top-5 words with the lowest distance between each other were selected. On these top words the labeling methods were applied.

### Finding labels with a scoring function

Trying to find a good label for topics we used topic words *w* and generated synsets *s* for each topic word with the help of WordNet. Based on them we picked their direct hypernyms *h*. To weight the hypernyms Custom scoring function (Csf) was defined, which includes the number of hypernyms *h* for the word *w* and the number of words, that have a hypernym in common. When a hypernym for a word was found the reciprocal of the total number of hypernyms for each word was assigned to to every hypernym of the current word. If a selected hypernym is used by another word, too, the scores are added and then multiplied by the number of common words. We select the final label by the highest score.

Figure 4.3 illustrates the scores for each hypernym, which are represented as circles above the hypernyms. The arrows connect the topic words *w* with their synsets *s*. These are connected to hypernyms *h*. For $w_1$ each hypernym $h_1, h_2$ has the value $\frac{1}{3}$. $h_4$ and $h_5$ have the value $\frac{1}{2}$, but $h_5$ is connected to $s_5$ and $s_6$. Therefore, we add up $\frac{1}{2}$ from $w_2$ and $\frac{1}{3}$ from $w_3$ and multiply the result by 2. In total $h_5$ reaches the highest score of $\frac{5}{3}$ and is selected as the final label.

### Find labels with similarity functions

The first one utilizes similarity functions provided by WordNet. The second one relies on Polyglot word embeddings to calculate the distance between two terms.

WordNet offers different similarity functions, to calculate the similarity between synsets:

- The *path-similarity* is defined by the nodes, which are visited while going from one word to another using the hypernym hierarchy. The distance between two words is the number of nodes that lie on the shortest path between two words in the hierarchy. The calculated score is in range of 0 and 1, while 1 means two words are identical.

- The *lch-similarity* (Leacock-Chodorow) is based on the shortest path $p$ and the maximum depth $d$ of the hierarchy in which the words occur. The path length is scaled by the maximum depth: $-log(p/2d)$

The remaining three similarity functions are measuring the Information Content (IC) of synsets. IC combines the knowledge of the hierarchical structure from WordNet, with statistics on actual usage in text as derived from a large corpus. Per default WordNet uses the Brown Corpus. Although, this corpus is not related to our domain-specific data, it includes a large number of English texts and is suitable as a reference corpus for this specific task.

- The *res-similarity* (Resnik-Similarity) weights edges between nodes by their frequency of the used textual corpus. Based on the IC of the Least Common Subsumer (lsc), the most specific ancestor node, a similarity score is calculated.

- The *jcn-similarity* (Jiang-Conrath Similarity) calculates the relationship between two words with $(IC(w_1) + IC(w_2) - 2 * IC(lcs))$ and

- the *lin-similarity* calculates it with $2 * IC(lcs)/(IC(w_1) + IC(w_2))$.

For all topic words we generated synsets and calculated for all possible combinations of the topic words the similarities of their synsets. For every possible topic word pair the highest similarity score from the synsets was taken and the lowest common hypernym was derived. If a combination of topic words had the same lowest common hypernym, the similarities were summed up. In the end, the hypernym with the highest score was taken as the final label.

The same procedure was applied also with Polyglot embeddings (plg). Instead of calculating the similarity between the synsets with WordNet similarity functions, the distance function from the Polyglot library was used. The lower the distance between two words the more similar they are. The other steps remained the same.

### 4.1.4 Evaluation

In the following section the results of intrinsic and extrinsic topic labeling will be evaluated regarding their quality and the number of different labels in a topic model. The labels generated automatically, are also compared to the manual labels, which were assigned by the domain experts. For the evaluation we used English editorial articles. First, we evaluate the intrinsic and second, the extrinsic topic labeling. Afterwards, the intrinsic and extrinsic labellings are compared with each other.

#### Intrinsic topic labeling

We applied the ATL in section 4.1.2 on our datasets, which include editorials, comments and forums. In general, the ATL according to Mei et al., 2007, outputs only different labels for topics, which were generated with LDA. For the topics generated with NMF the same label was given for every topic in a topic model. The reason could be, that NMF does not return a probability distribution for every document. Normalizing the values between 0 and 1 did not lead to an improvement. Therefore, the labels for topics generated with NMF were neglected. Further evaluations are based on English editorial articles.

**Topics from Generation 1** First, we used the topics from *Generation 1*, which were generated as described in 3.2. In Figure 4.4 the label counts for English editorial articles are shown. On the x-axis all labels are listen, while the y-axis denotes the number of topics, that the label was assigned to. Considering just the labels without verifying the topics, they are assigned to, the labels seem to be meaningfull and specific. Often, a label is a persons name e.g *Jose Andres, Rahm Emanuel, Morgan Stanly,Gloria Casas, Theresa Eisemann etc.*.

In Table 4.1 example topics are shown, which were labeled manually by domain experts and with the intrinsic approach. The intrinsic labels do not fit to the given topic: *Rahm Emanuel* an American politician is assigned to Topic 107, which deals with environment and waste. *Hairy vetch*, a plant variety, for Topic 23. *Irritable bowel syndrome* to Topic 64 and *Safran Foer*, an American novelist, to Topic 74, dealing with animal husbandry. The automatic labels have nothing in common with the manual ones.

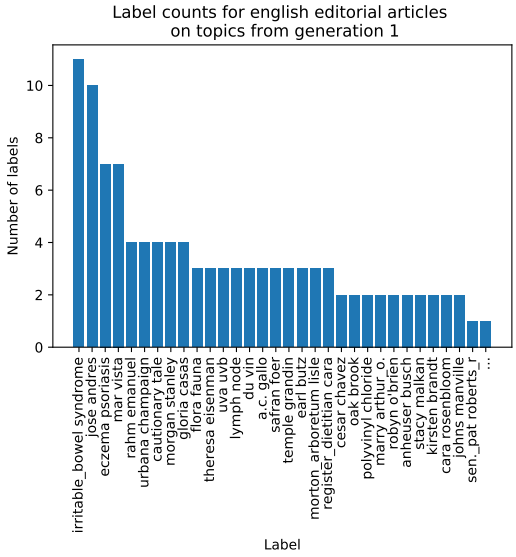



**Fig. 4.4.:** Label counts for topics from Generation 1 according to Mei et al., 2007.

| | Topic 107 | Topic 23 |
|---|---|---|
| method | waste, compost, use, scrap, material, landfill, ton, environmental, throw, gas | grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming |
| intrinsic<br>manual | rahm emanuel<br>waste | hairy vetch<br>homegrown food |
| | Topic 64 | Topic 74 |
| method | milk, raw, dairy, product, cheese,claim, health, cow drink, study | meat, feed, beef, animal, grass, cow, eat, raise, buy, make |
| intrinsic<br>manual | irritable bowel syndrome<br>dairy product | safran foer<br>animal husbandry |

**Tab. 4.1.:** Topics labeled manually and with intrinsic methods.

**Topics including POS-tagging:** By providing POS-tags, using Spacy[5], we can limit the labels to certain word types. In our experiments we used (NN-NN) or (JJ-NN) POS-tags for English topic labels and (NN-NN) or (ADJD-NN) for German. To apply POS-tagging, the preprocessing for the texts had to be changed, because in Generation 1, a collocation finder was used. After performing this step the POS-tags could not be applied retroactively. Therefore, we removed collocation finding and added POS-tagging. All other preprocessing steps remained the same. Nevertheless, the topics differ from the ones of Generation 1.

In Table 4.2 topics and labels are shown with different POS-tags. In comparison to the labels generated without POS-tagging, these labels seem closer to a topic. For Topic 6, 10, 23 and 37 the labels *music festival, premature aging, hunted games* and *modified organism* seem good.

[5]Possible POS-tags: https://spacy.io/api/annotation

| | Topic 6 | Topic 10 |
|---|---|---|
| with POS-tags | restaurant, fast, chain, meal, say, menu, ingredient, burger, chipotle, mcdonald | child, eat, kid, parent, family, healthy, school, who, health,can |
| (NN, NN) | music festival | anorexia nervosa |
| (JJ, NN) | hot fudge | premature aging |
| - | dunkin donuts | anorexia nervosa |
| | Topic 23 | Topic 37 |
| with POS-tags | meat, beef, feed,animal, grass, cattle,eat, raise, more, pork | carbon, climate, gas, greenhouse, emission, change, reduce, global, industrial, co2 |
| (NN, NN) | sport utility | gene splicing |
| (JJ, NN) | hunted game | interactive map |
| - | earl butz | modified organisms |

**Tab. 4.2.:** Labeled topics with intrinsic method

In Figure 4.5 the label counts for English editorial articles using the texts, that were POS-tagged are shown. On the x-axis all labels are listed, while the y-axis denotes the number of same labels. In the plots where POS-tags were applied, no labels include a name of persons and a smaller number of labels was outputted in contrast to the plot without POS-tags.

However, the same observation can be made as above. Although, the labels seem meaningful and specific they do not really fit to the topics. We assume that the high quality of the labels themselves stem from the way they are generated. By applying bi-gram mining on the original corpus only useful word combinations are found as candidate labels. That the labels seemingly do not fit to the topics means that measuring the relatedness between the topics and the labels by their KL-divergence is not successful on our data.

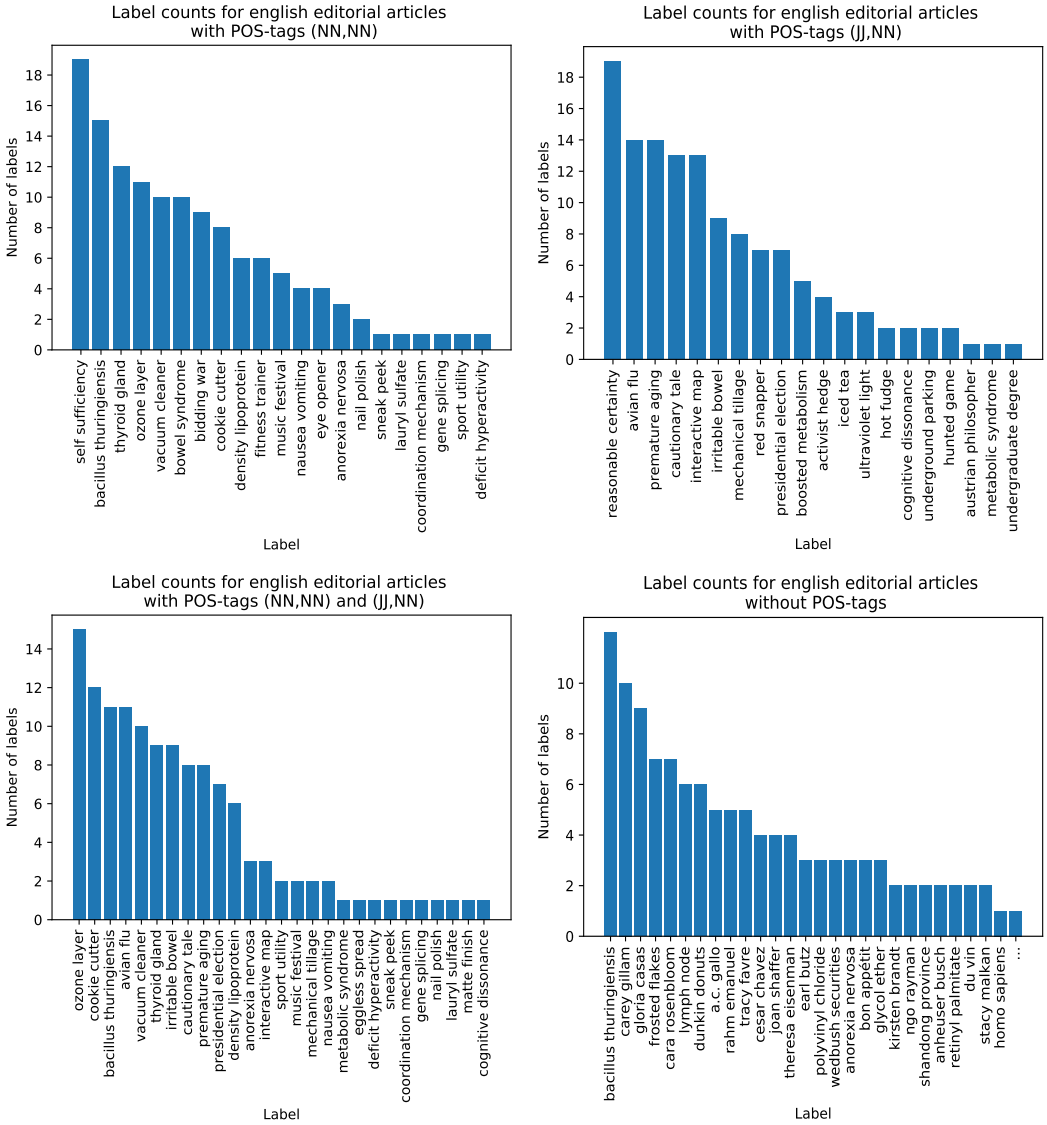



**Fig. 4.5.:** Label counts for topics including POS-tags with intrinsic method.

### Extrinsic topic labeling

Furthermore, we applied the ATL in section 4.1.3 on our Dataset, using the English online database WordNet and Polyglot embeddings. The described different similarity functions from WordNet, the Csf and the Polyglot embeddings were used to label our topics. A few examples are shown in Table 4.3 including the manual assigned labels to the topics, too. Some labels generated with the automatic approaches match the manual assigned labels. This is the case for the Topics 64, 84 and 107. For the other topics, the labels are heading to the same direction as the manual label: for Topic 97 *chemical* and manually *pesticide residues*, for Topic 99 *bee* and manually *beekeeping* and for Topic 109 *grocery store, mercantile establishment, marketplace* and manually *retailers* were assigned. Evaluating the automatically generated labels

| | Topic 23 | | Topic 64 | |
|---|---|---|---|---|
| method | grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming | | milk, raw, dairy, product, cheese, cow health, drink, study, claim | |
| path | entity | produce | abstraction | beverage |
| ich | entity | produce | abstraction | produce |
| res | produce | produce | **dairy product** | beverage |
| jsn | produce | produce | produce | beverage |
| lin | produce | produce | beverage | beverage |
| plg | vegetable | vegetable | **dairy product** | abstraction |
| Csf | cultivate | cultivate | nakedness | farm |
| manual | homegrown food | | **dairy product** | |
| | Topic 74 | | Topic 84 | |
| method | meat, feed, beef, grass, eat, raise, cow, buy, make, animal | | company, tea, brand, product, drink, honest, new, beverage, consumer, goldman | |
| path | entity | meat | **beverage** | **beverage** |
| ich | entity | abstraction | physical entity | substance |
| res | matter | meat | substance | substance |
| jsn | food | meat | **beverage** | **beverage** |
| lin | matter | meat | **beverage** | **beverage** |
| plg | cattle | physical entity | food | food |
| Csf | cattle | be | **beverage** | **beverage** |
| manual | animal husbandry | | **beverage** | |
| | Topic 97 | | Topic 99 | |
| method | fruit, vegetable, pesticide, produce, buy, eat, list, apple, residue, sweet | | bee, honey, study, hive, year, beekeeper, plant, researcher, honeybee, colony | |
| path | matter | matter | organism | person |
| ich | matter | matter | organism | person |
| res | matter | matter | organism | organism |
| jsn | matter | matter | organism | whole |
| lin | produce | matter | bee | artifact |
| plg | fruit | entity | bee | artifact |
| Csf | chemical | chemical | farmer | scientist |
| manual | pesticide residues | | beekeeping | |
| | Topic 107 | | Topic 109 | |
| method | waste, compost, use, scrap, material, landfill, ton, environmental, throw, gas | | foods, company, store, chain, market, executive, new, year, mackey, grocery | |
| path | material | material | grocery store | mercantile establishment |
| ich | abstraction | physical entity | physic entity | mercantile establishment |
| res | material | material | social group | mercantile establishment |
| jsn | abstraction | material | grocery store | mercantile establishment |
| lin | material | material | social group | mercantile establishment |
| plg | **waste** | abstraction | artifact | abstraction |
| Csf | convent | lowland | marketplace | marketplace |
| manual | **waste** | | retailer | |

**Tab. 4.3.:** Topics labeled from Generation 1 manually and with extrinsic methods. Labels including preprocessing are in the third and fifth column. **Bold** words are the same as the manual assigned label.

| method | entity | physical entity | object | whole | matter | abstraction | $\sum$ |
|---|---|---|---|---|---|---|---|
| path | 19 | 20 | 7 | 4 | 1 | 33 | 84 |
| | **7** | **7** | **5** | **2** | **1** | **16** | 38 |
| ich | 29 | 23 | 7 | 4 | 1 | 42 | 106 |
| | **13** | **13** | **9** | **3** | **1** | **25** | 64 |
| res | - | 4 | 5 | 4 | 9 | 5 | 27 |
| | - | **2** | **4** | **1** | **2** | **1** | 10 |
| jsn | 19 | 14 | 3 | 2 | 1 | 25 | 64 |
| | 10 | **6** | **2** | **2** | **2** | **9** | 31 |
| lin | - | 1 | 8 | 6 | 9 | 11 | 35 |
| | - | **1** | **3** | **5** | **3** | **5** | 17 |
| plg | 1 | 1 | 3 | 6 | 4 | 3 | 18 |
| | **7** | **7** | **4** | **7** | **3** | **19** | 47 |

**Tab. 4.4.:** Label counts of non informative words with different similarity functions. **Bold** numbers denote labels including preprocessing.

using different approaches, it was discovered that depending on the topics different labeling techniques output the best labels. It is not possible to tell, which approach is the best for all topics, let alone for several topic models according to the labels. Therefore, we tried to evaluate the labels generated with the extrinsic methods according to label counts. The words *entity, physical entity, object, whole, matter* and *abstraction* were chosen, because these are the most general words in the hierarchical tree of hypernyms in WordNet and do not have a high informative value. In Table 4.4 the number of non informative words are listed for the different similarity functions from WordNet. Based on the sum of the non informative words per similarity function and Polyglot embeddings (plg), we ranked the different methods in Table 4.5. The top 3 are: res-similarity with preprocessing, lin-similarity with preprocessing and Polyglot embeddings. The labels with Csf do not include

| | | |
|---|---|---|
| 1. **res-similarity** | 2. **lin-similarity** | 3. polyglot embeddings (plg) |
| 4. res-similarity | 5. **jsn-similarity** | 6. lin-similarity |
| 7. **path-similarity** | 8. **polyglot embeddings** | 9. jsn-similarity |
| 10. **ich-similarity** | 11. path-similarity | 12. ich-similarity |

**Tab. 4.5.:** Ranked similarity functions. **Bold** similarities denote the similarities, which were applied on preprocessed topics.

any non informative words, because only the direct hypernyms and not the whole hierarchy of hypernyms were considered. Therefore, we plotted the amount of distinct labels in Figure 4.6. This shows, the labels generated with preprocessing on the left side and the labels without on the right. The number of same labels is at most 6 or 8, which shows that the labels are discriminative.

Having evaluated the intrinsic and extrinsic automatic topic labeling we can conclude, that the intrinsic approach generates meaningful and specific labels, that do not

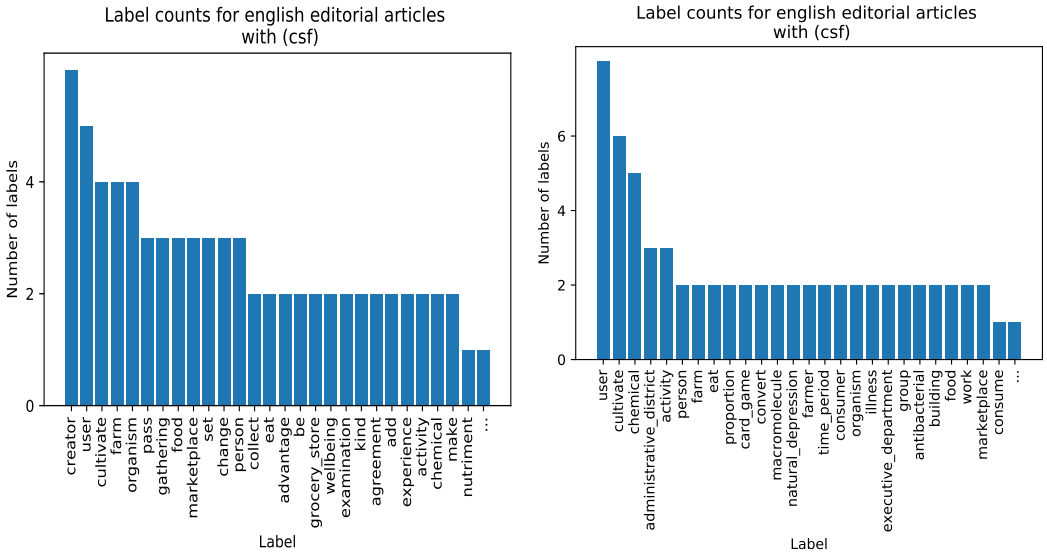



**Fig. 4.6.:** Label counts for topics from Generation 1 with Csf.

fit to the topics. The extrinsic approach generates partially good results, which are comparable with the labels from the domain experts. Nevertheless, finding meaningful and high qualitative labels is not yet automatable. The knowledge and experience a human person, which is required for topic labeling, can not be replaced by a machine.

## 4.2 Internal consistency

When generating a topic model with LDA or NMF the number of topics has to be manually set. This number is critical and has an effect on the quality and the interpretability of topics. We want to provide the domain experts an overview how topics change when increasing or decreasing the topic number. So they can assess, which topic model is the appropriate one for their further research.

To analyze the quality of a topic model we differentiate between the intra topic model and inter topic model approach. The intra topic approach compares all topics of one topic model with each other. This way we can study, which topics are similar to each other or which topics appear together in the same document. When applying the inter topic approach, we compare the topics from topic model A to a second different topic model B. The second topic model differs from the first by the number of topics. By comparing the two models we can study what effect the increasing topic number has on the quality of the topics. With both approaches we want to examine questions such as: Do the topics get more specific, more general, do they split up or do they stay the same and only new topics are added? Are there a few topics, which are dominating in a document or are few topics assigned to a document? How does the topic assignment change across different topic models? Indicates a higher topic number a better clustering of the documents?

Both, LDA and NMF return a document topic matrix $\theta$, which describes to what extend a topic appears in a document and a topic term matrix $\phi$, which describes to what extend a term appears in a topic. Different key figures can be derived from these matrices to judge the quality and to examine the changes in topics. Entropy and Jensen Shannon divergence can be used on the document topic matrix as well on the topic term matrix. The coherence measure relies solely on the topic term matrix and alpha $\alpha$ can only be applied on topic models, that were generated with LDA.

For our evaluation we generated topic models with 25, 50, 75 topics and topic models with 50 topics over and under the topics number from *Generation 1*. The different key figures were applied on these newly generated topic models and the topic models from *Generation 1*.In the following the dataset for German editorial articles with 25, 50, 75, 140, 190 and 240 topics per topic model were analyzed.

| | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| Document 1 | 0.1 | 0.4 | 0.05 | 0.25 | 0.2 |
| Document 2 | 0.025 | 0.8 | 0.025 | 0.07 | 0.03 |

**Tab. 4.6.:** Example for a document topic matrix

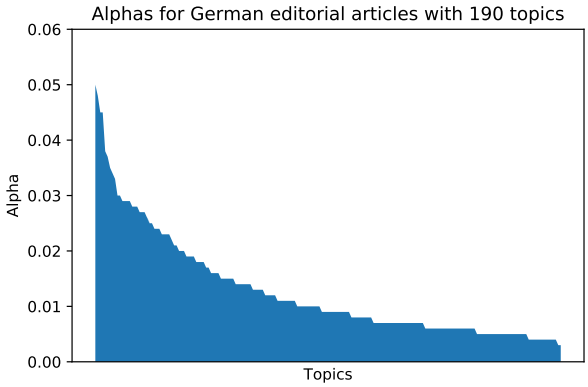
### 4.2.1 Theta $\theta$

The document topic matrix $\theta$ describes to which extend a topic is represented in a certain document. We used the matrix to calculate the number of documents, which are assigned to a topic and the number of topics, which are assigned to a document. In both cases a threshold fo 10% was used. The example document topic matrix in Table 4.6 shows two documents and 5 topics. For topic 1 the counter number of documents is only 1 (Document 1). For Document 1 the relevant number of topics is 4 (Topic 1,2,4,5).

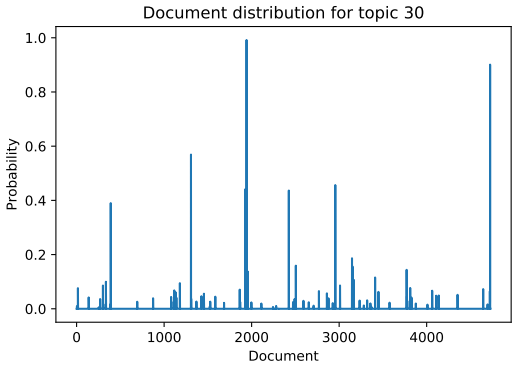
### 4.2.2 Alpha

Alpha $\alpha$ is a prior parameter for LDA, that describes the sparsity of the topic distribution for every document. Usually, the prior $\alpha$ has to be set and has the same value for every topic. In this case it is called the symmetric Dirichlet prior. However, Blei, 2003 showed how $\alpha$ can be estimated from the data per topic. In this case $\alpha$ is an asymmetric Dirichlet prior. This method was used for our topic model, to determine how important the topics are for the whole corpus. A high $\alpha$ value means that the documents are mixtures of many topics, while a low $\alpha$ value means, that the documents are composed of only a few highly probable topics (Steyvers and Griffiths, 2007a).

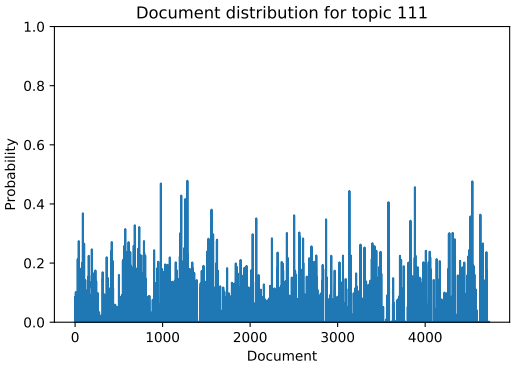
In Figure 4.7a the alpha values for each topic for German editorial articles with 190 topics are shown. The document distribution for Topic 30, which has the lowest alpha value, is considered in Figure 4.7b. The x-axis represents the document ids, standing for the documents, which build the corpus. The y-axis represents the percentage of a document that is covered by the topic. In contrast to Topic 30, Topic 111 with the highest alpha value is shown in Figure 4.7c. One can see, that Topic 30 covers only a few documents, while Topic 111 is more evenly spread over all documents.




**(a)** Plotted alphas for German editorial articles




**(b)** Document coverage for the topic with the lowest alpha value




**(c)** Document coverage for the topic with the highest alpha value

**Fig. 4.7.:** Alpha values for German editorial articles with 190 topics and the topic document matrices for the topic with the highest and lowest alpha value

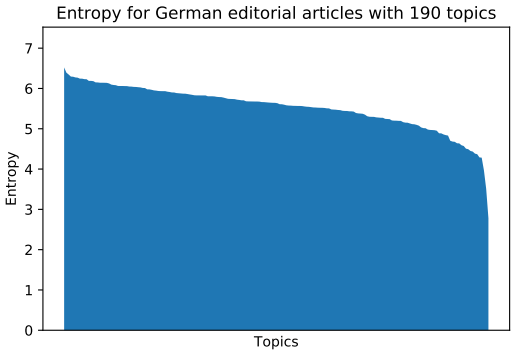
### 4.2.3 Entropy

Entropy was used to identify specific and general topics. It can be applied on the topic term matrix $\phi$ and the document topic matrix $\theta$. When applied on the topic term matrix, a high entropy value indicates that the topic is rather general. This means, that all terms have a similar probability to appear in the topic. A low entropy indicates, that the topic is specific i.r. only a few words have a high probability to appear in the topic. This difference is illustrated in Figure 4.8b and in Figure 4.8c. When applied on the document topic matrix $\theta$ the rules can be applied analogously. A high entropy value indicates, that the topic is rather general. This means, that all topics have a similar probability to appear in a document. A low entropy value indicates, that the topic is specific i.e. only a few topics have a high probability to appear in a document. Entropy is calculated as follows (Ankit Sethi, 2012):

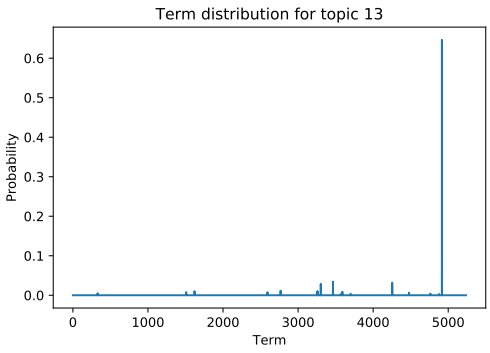
$$E = -\sum(p * log(p)) \tag{4.4}$$

where $p$ is the probability of a term in a topic, which was taken from the topic term matrix $\phi$ or the probability of a topic in a document, taken from the document topic

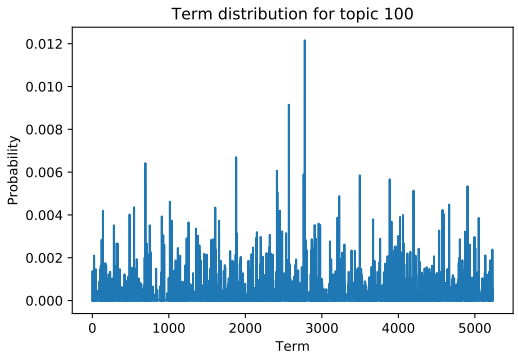
matrix $\theta$. In Figure 4.8a the entropy values for each topic for German editorial articles with 190 topics are shown. The entropy values are sorted descending. The topic term matrix for Topic 13, which has the lowest entropy value, is considered in Figure 4.8b. The x-axis represents each termid of the corpus. The y-axis represents the probability of the terms in a topic. In contrast to Topic 13, Topic 100 with the highest entropy value is shown in Figure 4.8c. One can see, that Topic 13 consists mainly of the term id 4916 and the other term ids hardly occur, while in Topic 100 the term ids are nearly evenly spread.




**(a)** Plotted entropy for German editorial articles




**(b)** Topic coverage for the topic with the lowest enropy




**(c)** Topic coverage for the topic with the highest entropy

**Fig. 4.8.:** Entropy for German editorial articles with 190 topics and the topic term matrices for the topic with the highest and lowest entropy

### 4.2.4 Coherence

The coherence scores topics by measuring the degree of semantic similarity between words in a topic. This measurement helps to distinguish topics, that are semantically similar and easy interpretable for humans and those that are semantically dissimilar and not easy interpretable.(Stevens, Keith;Kegelmeyer,Philip;Andrzejewski, David;Buttler, 2012) There are different coherence measure such as the UCI-measure (Newman et al., 2010) and the U-mass measure(Mimno et al., 2011). Both measure

the coherence of a topic as the sum of pairwise distributional similarity scores over a set of topic words:

$$coherence(V) = \sum_{v_i, v_j \in V} score(v_i, v_j) \tag{4.5}$$

$V$ describes the set of words for a topic, while $v$ is a single word, occurring in a topic. We used the top-10 words of a topic to calculate the coherence score for each topic in a topic model. We used the UMass metric, which is based on the document co-occurrence and defined as:

$$socre(v_i, v_j) = log \frac{D(v_i, v_j) + 1}{D(v_j)} \tag{4.6}$$

$D(v_j)$ is the document frequency, that count the number of documents which include the word $v_j$. $D(v_i, v_j)$ is the co-document frequency, that counts the documents, which include both word $v_i$ and $v_j$. A smoothing count of 1 is included to avoid taking the logarithm of zero. The UMass metric computes these counts over the original corpus, which was used to train the topic models (Stevens, Keith;Kegelmeyer,Philip;Andrzejewski, David;Buttler, 2012). The coherence score is negative and the higher interpretability of a topic is given with a score near to zero.

### 4.2.5 Jensen Shannon divergence

The topics returned by LDA are probability distributions over all terms in the corpus. Therefore, to compare the similarity of two topics $p$ and $q$ we can use existing metrics to measure the similarity between probability distribution. Lin, 1991 et al lists possible similarity functions. A standard function to measure the difference or divergence between two probability distributions $p$ and $q$ is the Kullback Leibler (KL) divergence:

$$D(p, q) = \sum_{j=1}^{T} p_j log_2 \frac{p_j}{q_j} \tag{4.7}$$

where $j$ is the number of a certain term and $T$ describes the total number of terms. $p_j$ represents the probability of term $j$ appearing in topic $p$. $q_j$ is the probability of term $j$ appearing in topic $q$. The KL divergence is an asymmetric measurement. For our use-case, comparing the topics intra and inter a topic model, a symmetric measure is needed, which guarantees the same results for the comparison of $t1$ with $t2$ and $t2$ with $t1$. Therefore, based on the KL divergence, the Jensen Shannon (JS) divergence is used:

$$JS(p, q) = 0.5 * (D(p, \frac{p+q}{2}) + D(q, \frac{p+q}{2})) \tag{4.8}$$

The JS divergence is a symmetric extension of the KL divergence. If the probability distributions are identical, the value 0 is assigned otherwise the value 1 is assigned for totally dissimilar probability distributions (Steyvers and Griffiths, 2007a). In our implementation we subtracted the value from the JS divergence from 1 to get the similarity between two probability distributions, so that the value 1 is assigned when two probability distributions are identical and the value 0 when they are completely dissimilar.

### 4.2.6 Evaluation

The evaluation was conducted on *German* and *English editorial articles*. For both datasets we studied topic model with 50 topics over and under the optimal topic number from *Generation 1* and further generated topic models with 25, 50 and 75 topics. Both datasets were analyzed by applying each key figure as explained above on each topic model. The results per key figure and per topic model were then compared with each other.

#### Entropy

The Figures 4.9 and 4.10 show the change of entropy for *German* respectively *English editorial articles*. The table on the left denotes the minimal and maximal entropy given the number of topics. On the right the maximal entropy values (blue line) and the minimal entropy values (orange line) are plotted. This structure will repeat in the different key figures.

The entropy values for *German editorial articles* (Figure 4.9) are decreasing when increasing the number of topics. For the topic model with 240 topics the entropy values starts increasing again. This means, that the topics with a higher topic number are getting more specific until a certain point, when the entropy value is increasing again. This could mean, that the optimal topic number, to generate topics, which are specific, is at the point, when the entropy value has reached its minimum.

For *English editorial articles* (Figure 4.10) the maximal entropy value is decreasing up to the topic model with 80 topics. Then the values is rising and for the topic models with 130 and 180 topics the entropy stays the same. For the minimal entropy the values are getting continuously smaller. So the span between the maximal and the minimal entropy value is increasing. This means, that the more topics are generated the more topics get more general and more specific.

| | max value | min value |
|---|---|---|
| 25 topics | 7.08 | 5.42 |
| 50 topics | 6.82 | 4.67 |
| 75 topics | 6.79 | 4.40 |
| 140 topics | 6.60 | 3.99 |
| 190 topics | 6.52 | 2.78 |
| 240 topics | 6.57 | 3.09 |

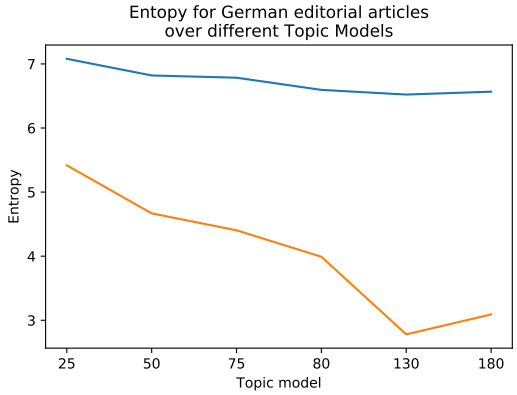



**Fig. 4.9.:** Maximal and minimal entropy per topic model for German editorial articles.

| | max value | min value |
|---|---|---|
| 25 topics | 7.029 | 5.455 |
| 50 topics | 6.8 | 4.71 |
| 75 topics | 6.823 | 4.636 |
| 80 topics | 6.819 | 4.112 |
| 130 topics | 8.741 | 4.005 |
| 180 topics | 8.741 | 3.249 |

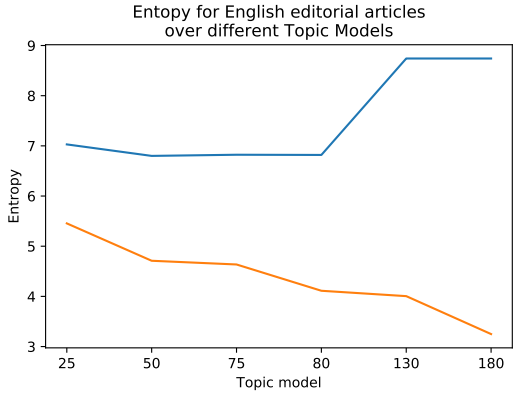



**Fig. 4.10.:** Maximal and minimal entropy values per topic model for English editorial articles.

### Alpha

In Figure 4.11 the alpha values for *German editorial articles* are shown. The maximal alpha values as well the minimal alpha values are decreasing. This indicates, that the documents are described by fewer topics with a higher probability. But alpha does not say anything about the topic quality, so the few topics, which are assigned to the document can be rather general or specific. Therefore, we calculated the entropy for the topic with the maximal alpha value 0.186 from the topic model with 25 topics and the minimal alpha value 0.002 from the topic model with 240 topics. We got the entropy value of 7.08 for the topic model with 25 topics and the entropy value of 6.28 for the topic model with 240 topics. So the document with the highest alpha value consists out of either general topics and the document with the lowest alpha value out of specific ones.

The minimal alpha values for *English editorial articles* (Figure 4.12) are continuously declining up to the topic model with 130 topics. Then the alpha value is staying the

same. The maximal alpha values are volatile, so there is no prediction possible, if the values will raise or fall again.

| | max value | min value |
|---|---|---|
| 25 topics | 0.186 | 0.027 |
| 50 topics | 0.104 | 0.015 |
| 75 topics | 0.065 | 0.011 |
| 140 topics | 0.061 | 0.004 |
| 190 topics | 0.050 | 0.003 |
| 240 topics | 0.046 | 0.002 |

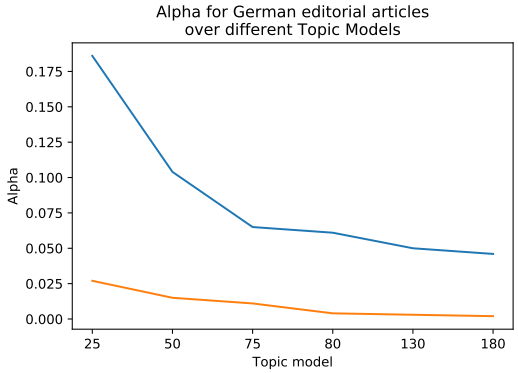



**Fig. 4.11.:** Maximal and minimal alpha values per topic model for German editorial articles.

| | max value | min value |
|---|---|---|
| 25 topics | 0.159 | 0.032 |
| 50 topics | 0.229 | 0.013 |
| 75 topics | 0.159 | 0.010 |
| 80 topics | 0.180 | 0.004 |
| 130 topics | 0.144 | 0.002 |
| 180 topics | 0.118 | 0.002 |

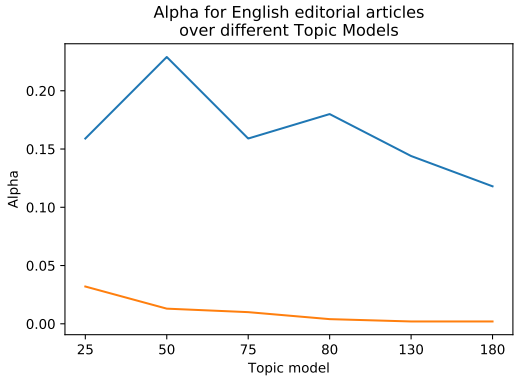



**Fig. 4.12.:** Maximal and minimal alpha values per topic model for English editorial articles.

### Coherence

The maximal coherence score is for *German editorial articles* (Figure 4.13) in the range of the maximal values -0.91 and -1.1 and for *English editorial articles* (Figure 4.14) in the range of -0.49 and -0.56. The maximal values do not change a lot for both datasets, but no pattern how the values are changing can be seen. The same can be said for the minimal coherence values, the only difference is, that the range in which the coherence is moving is much bigger than the range from the maximal values. For *German editorials* it is between -3.96 and -9.2 and for *English editorials* between -2.2 and -16.4.

| | max value | min value |
|---|---|---|
| 25 topics | -0.994 | -3.957 |
| 50 topics | -0.999 | -5.605 |
| 75 topics | -1.102 | -6.792 |
| 140 topics | -0.956 | -6.661 |
| 190 topics | -0.912 | -9.19 |
| 240 topics | -0.948 | -7.25 |

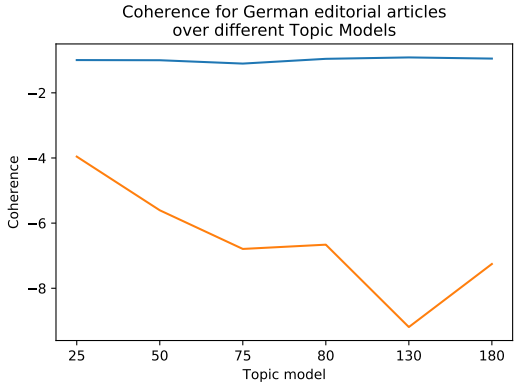



**Fig. 4.13.:** Maximal and minimal coherence values per topic model for German editorial articles.

| | max value | min value |
|---|---|---|
| 25 topics | -0.557 | -2.242 |
| 50 topics | -0.568 | -9.570 |
| 75 topics | -0.505 | -5.913 |
| 80 topics | -0.493 | -3.666 |
| 130 topics | -0.525 | -16.257 |
| 180 topics | -0.549 | -16.374 |

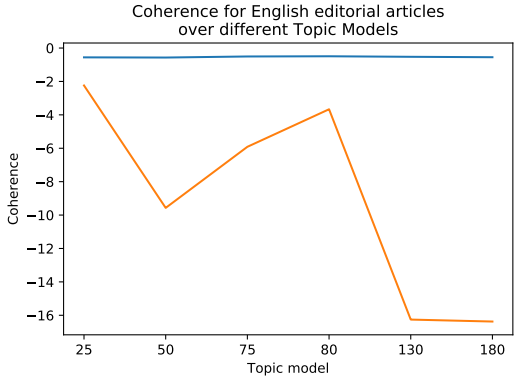



**Fig. 4.14.:** Maximal and minimal coherence values per topic model for English editorial articles.

### Theta

In the following we used the document topic matrix to calculate, in how many documents a certain topic covers more then 10% of the document and how many topics occur over 10% in a document.

First, we start with the number of documents a certain topic covers over 10%. In Figure 4.15 one can see, that the maximal and minimal number of documents is decreasing when increasing the topic number for *German editorial articles*. This indicated, that the topics are so specific, that they do not appear in any document with a probability higher then the threshold.

For *English editorial articles* (Figure 4.16) the maximal number of documents is volatile, whereas the minimal numbers of documents are decreasing and there even seem to be topics, that are not assigned to any document, because they do not cover the meaning of the document over the threshold.

| | max value | min value |
|---|---|---|
| 25 topics | 1878 | 154 |
| 50 topics | 1026 | 46 |
| 75 topics | 693 | 32 |
| 140 topics | 655 | 11 |
| 190 topics | 423 | 5 |
| 240 topics | 352 | 5 |

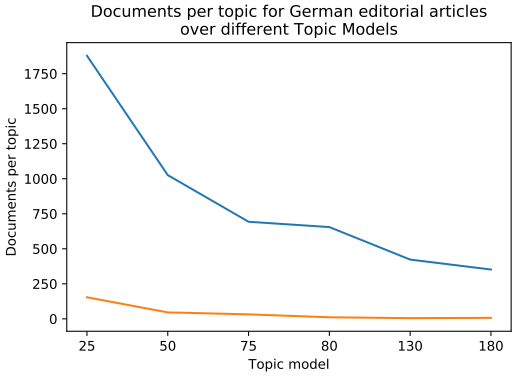



**Fig. 4.15.:** Maximal and minimal number of documents containing the same topic for German editorial articles.

| | max value | min value |
|---|---|---|
| 25 topics | 788 | 58 |
| 50 topics | 589 | 13 |
| 75 topics | 435 | 7 |
| 80 topics | 636 | 7 |
| 130 topics | 292 | 0 |
| 180 topics | 443 | 0 |

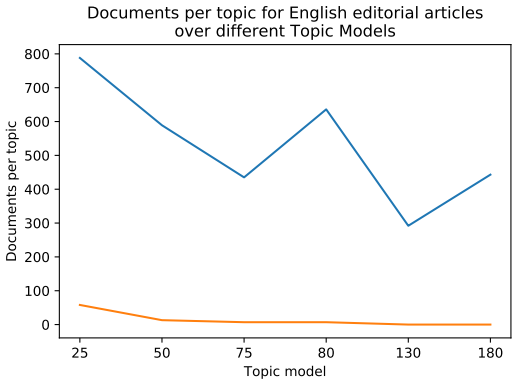



**Fig. 4.16.:** Maximal and minimal number of documents containing the same topic for English editorial articles.

Second, we analyzed how many topics occur in a document over 10%. The plots in Figure 4.18 and 4.17 represent the number of documents with the number of topics over the threshold. On the x-axis the amount of topics, which are occurring over 10%in a document, while the y-axis represents the number of documents, which include a certain amount of topics.

In Figure 4.17 is shown how the number of topics for *English editorial articles* change. When increasing the number from 25 to 50 topics, all documents are modeled by a most 6 topics. This means, that with a higher topic number the topics are more tailored to documents and thus the documents can be represented with less topics. This is also supported by the observation, that the number of documents, that contain only one or two topics over the threshold are slightly increasing with the number of topics. At the same time there are some documents, that do not express any topic over the threshold, which could indicate, that these documents are rather general and cover many topics with a low probability, instead of covering a few specific topics. For the *German editorial articles* in Figure 4.18 it can be seen, that the maximal number of topics, covering a document is also at 3. From the topic model with 25 topics to 50 topics, there are no documents, which include 8 topics, but

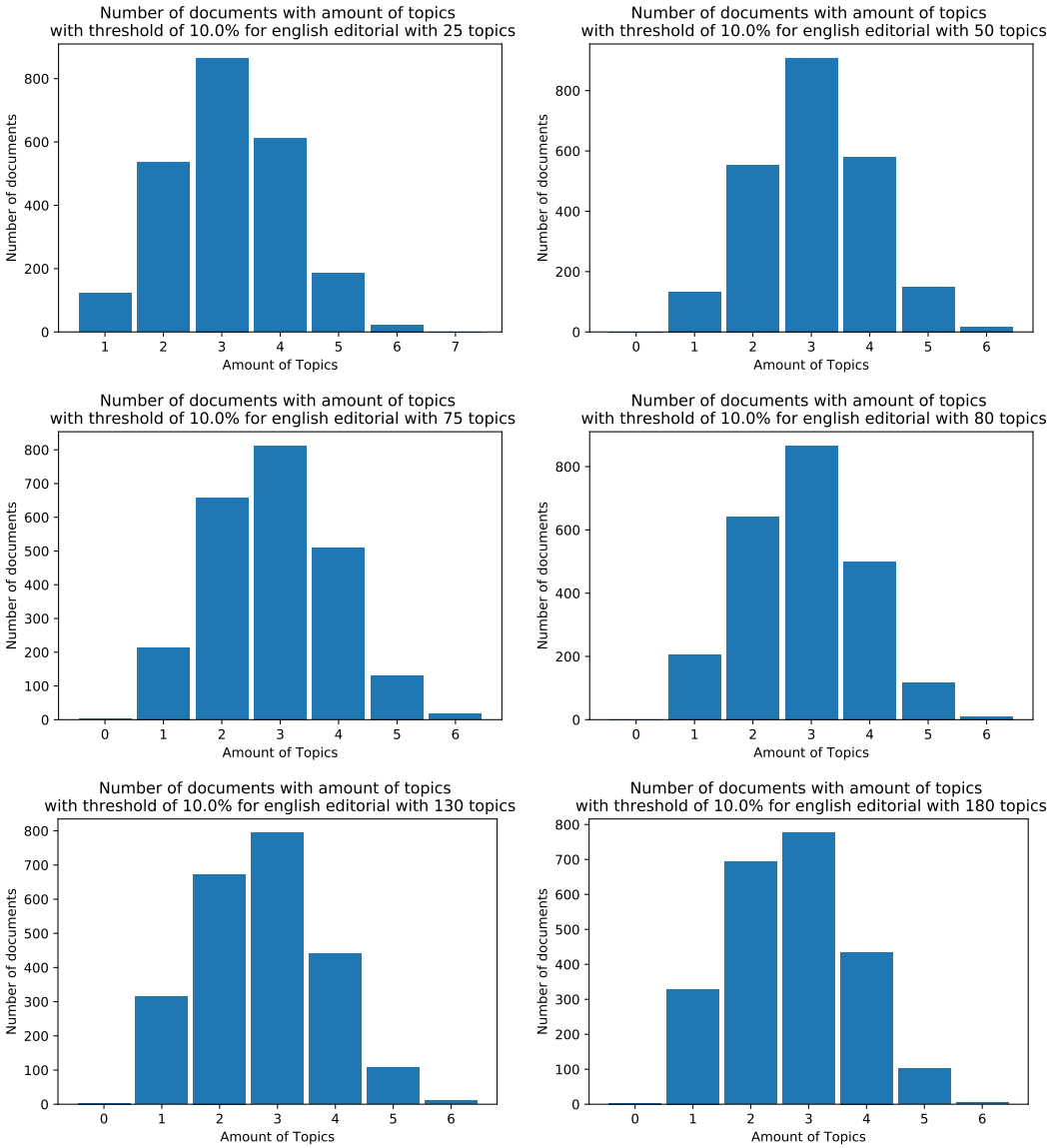



**Fig. 4.17.:** Amount of topics in documents over a threshold of 10% for English editorial articles

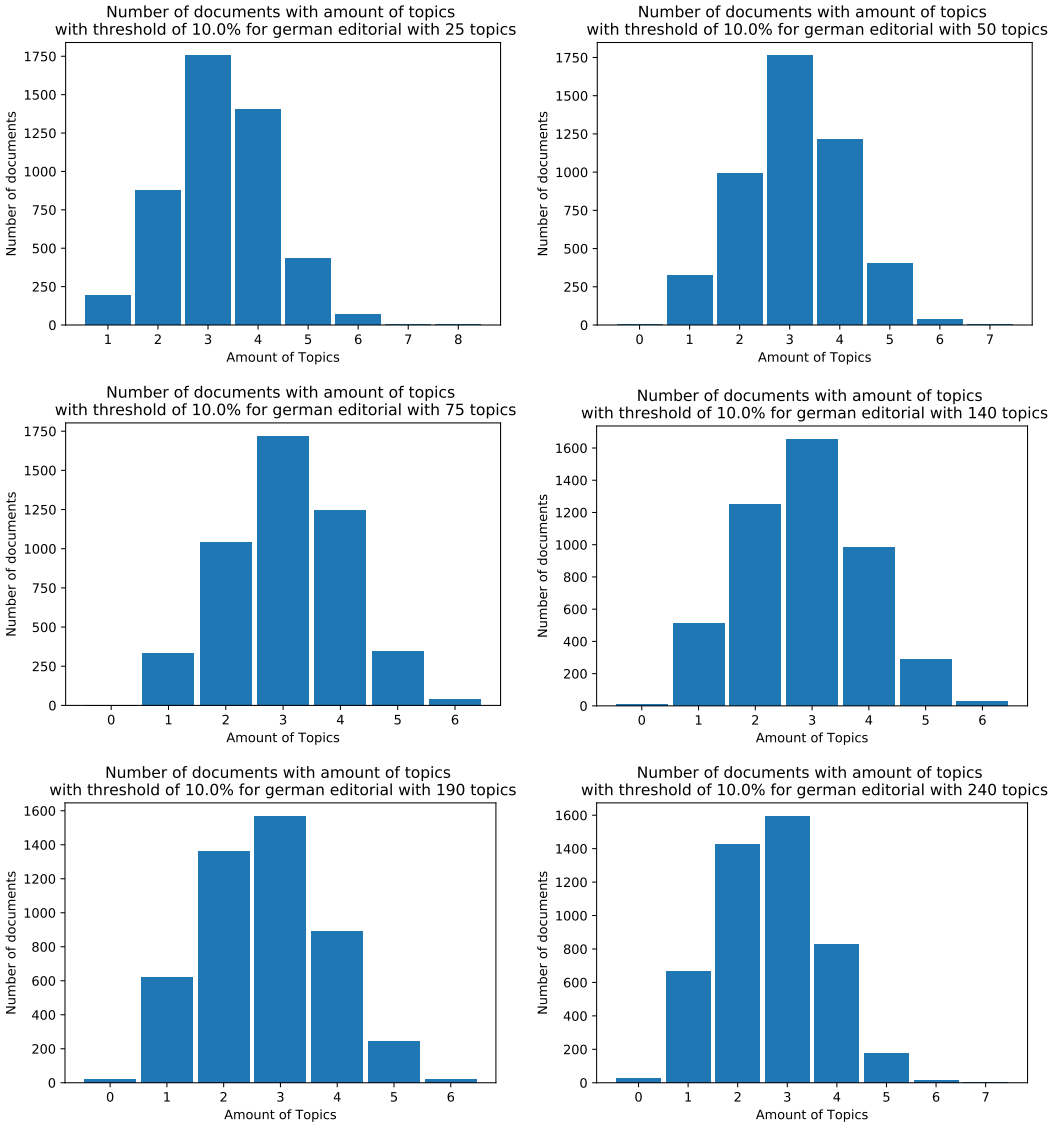



**Fig. 4.18.:** Amount of topics in documents over a threshold of 10% for German editorial articles

there are documents, that do not contain any topic. When increasing the number of topics to the next level, there are no documents, which contain 7 topics. In the next plots, the number of documents containing 2, only 1 or no topic is raising, but the maximal topic number remains 3. The last plot is looking nearly the same apart from adding a few documents with 7 topics. For this dataset it could be said, that the topics are getting up to 190 topics more specific. With more topics the number of generic topics increases. So the topic model with 190 seems the best for obtaining good topics.

### Jensen Shannon divergence

The JS divergence was used to calculate the similarity of topics based on the topic term matrix in a topic model and across topic models. The similarities were then visualized as heat maps. The dark red colored boxes express high similarity, while the brighter boxes express a smaller similarity. The topic models form *German editorial articles* with 25, 50 and 70 topics were evaluated manually. The highest similarities (dark red) with the values of 0.63, 0.65, and 0.65 for the topics intra a topic model were taken, and the topics were compared with each other in Table 4.7. The same words form the top 10 terms of a topic were marked in bold. In Table 4.8 the similarities of the topics were calculated across different topic models with 25/50, 50/57 and 50/75 topics. The values for the similarity were 0.9, 0.87 and 0.94. One can see, that the number of common words and the calculated similarity, is much higher in the inter topic model evaluation. This means, that the topics inter a topic model are not as similar to each other as the topics developing across topic models. At least this is not recognizable by the top 10 words of a topic. The heat maps for the inter topic model and intra topic model evaluation can be seen in Figure B.13 and B.10.

Within the different topic models the topics are hardly similar. This is a good characteristic, because it shows, that the topic number was not over fitted. Across the topic models there are different topics, but also very similar topics, which is shown in Table 4.19 for *German editorial articles* and in Table 4.20 for *English editorial articles*. This means, when increasing the number of topics, some of the new generated topics do not change much and cover the same themes such as the previous topic model, but new topics are added, too, which cover new themes. By adding new topics the development of the topics can be tracked.

| topics | compared topics with the highest similarity: | |
|---|---|---|
| 25 | T1: all, jed, sehen, stehen, leben, einfach, welt, finden, frage, bio | T19: bauer, landwirt, landwirtschaft, milch, preisen, kuh, betrieb, hof, euro, cent |
| 50 | T42: essen, lebensmittel, **jed**, fleich, all, kaufen, leben, **stehen**, ernährung, einfach | T40: hof, betreiben, landwirtschaft, **stehen**, landwirt, bauer, familie, verkaufen, **jed**, alt |
| 75 | T63: bauer, landwirt, euro, **preisen**, konventionell, biobauer, geld, bekommen, umstellen, ernten | T67: product, verbraucher, kunde, deutschland, **preisen**, deutsch, prozent, handeln,supermarkt, deutsche |

**Tab. 4.7.:** Compared topics intra a topic model with the highest similarity. Common topic words are **bold**

| topics | compared topics with the highest similarity: | |
|---|---|---|
| 25/50 | T21: **prozent**, **euro**, **ökologisch**, **betrieb**, **hektar**, million, steigen, **fläche**, deutschland, **zahlen** | T2: **prozent**, **ökologisch**, **betrieb**, **hektar**, **fläche**, landwirtschaft, **zahlen**, anteil, bewirtschaften, **euro** |
| 50/75 | T4: **pestizid**, **finden**, **probe**, **rückstand**, **greenpeace**, konventionell, **untersuchen**, **belasten**, prozent,einsatz | T54: **pestizid**, **rückstand**, **probe**, grenzwert, **finden**, **greenpace**, stoff, **belasten**, **untersuchen**, einsetzen |
| 50/75 | T9: **eiern**, **fipronil**, **belasten**, **niederlande**, **deutschland**, **nehmen**, **verkaufen**, **betreffen**, betrieb, angeben | T57: **eiern**, **fipronil**, **belasten**, **niederlande**, **nehmen**, **deutschland**, **verkaufen**, betroffen, **betreffen**, behörde |

**Tab. 4.8.:** Compared topics inter topic models with the highest similarity. Common topic words are **bold**

| topics | min value | max value |
|---|---|---|
| 25 | 0.38 | 0.63 |
| 50 | 0.35 | 0.65 |
| 75 | 0.33 | 0.65 |
| 140 | 0.34 | 0.62 |
| 190 | 0.33 | 0.61 |
| 240 | 0.33 | 0.60 |

**(a)** Minimal and maximal similarities for German editorial articles inter different topic models

| topics | min value | max value |
|---|---|---|
| 25/50 | 0.35 | 0.91 |
| 25/75 | 0.35 | 0.88 |
| 50/75 | 0.35 | 0.94 |
| 140/190 | 0.33 | 0.88 |
| 140/240 | 0.32 | 0.92 |
| 190/240 | 0.33 | 0.91 |

**(b)** Minimal and maximal similarities for German editorial articles intra different topic models

**Fig. 4.19.:** Minimal and maximal similarities inter and intra topic models for German editorial articles

| topics | min value | max value |
|---|---|---|
| 25 | 0.42 | 0.66 |
| 50 | 0.36 | 0.68 |
| 75 | 0.35 | 0.69 |
| 80 | 0.35 | 0.67 |
| 130 | 0.34 | 0.65 |
| 180 | 0.36 | 0.67 |

**(a)** Minimal and maximal similarities for English editorial articles inter different topic models

| topics | min value | max value |
|---|---|---|
| 25/50 | 0.39 | 0.90 |
| 25/75 | 0.37 | 0.91 |
| 50/75 | 0.35 | 0.91 |
| 80/130 | 0.34 | 0.95 |
| 80/180 | 0.33 | 0.97 |
| 130/180 | 0.33 | 1.0 |

**(b)** Minimal and maximal similarities for English editorial articles intra different topic models

**Fig. 4.20.:** Minimal and maximal similarities inter and intra topic models for English Editorial articles

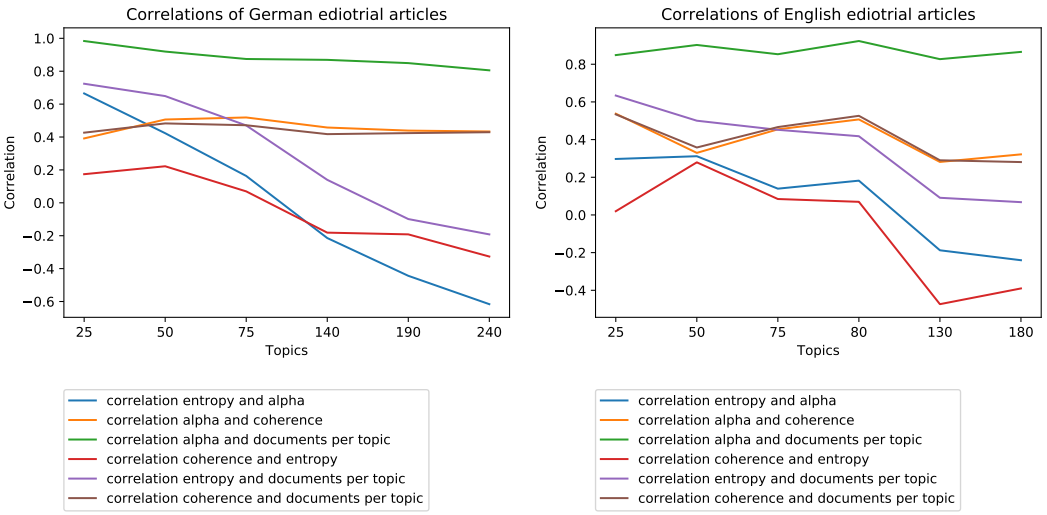
### Correlation

Calculating the correlation between every key figure, the relationship over the different topic models is analyzed by using the Pearson correlation. It is a measure of the linear correlation between two variables $X$ and $Y$ and can assume values between 1 an -1, where 1 represents a total positive linear correlation, 0 no linear correlation and -1 represents a total negative linear correlation.

In Figure 4.21 the correlation for the key figures alpha, entropy, coherence and documents per topic is calculated. On the x-axis the topic models with different topics is shown, while the y-axis shows the calculated correlation.

One can see, that the alpha and documents per topic values are correlating. This is, because both measure the sparsity of a topic over documents in a corpus. The alpha values, coherence values and documents per topic values hardly correlate. From this one can deduce, that the topics, which are most present in a document are not so easily interpretable for humans. Furthermore, all correlations including entropy are at the beginning highly correlating, but the correlation coefficient is continuously decreasing. So entropy in general seem to be a metric, which is developing independently form the other key figures, so from the characteristic, if a topic is rather general or specific, can not be deduced, if the topic is strongly represented in a document or if it is easily interpretable.

The analysis of the different key figures on the two datasets has shown that no key figure on its own is enough to determine the quality or the optimal topic number. When considering multiple key figures, it is possible to compare different topic models based on how generic or specific the topics are (entropy) or how the topics are distributed across the corpus (alpha). It is not possible, however, to compare topic models of different datasets. Each dataset has its own characteristics such as the number of documents, the average length of the documents or how specific the




**(a)** Correlation for German editorial articles **(b)** Correlation for English editorial articles

**Fig. 4.21.:** Correlations of key figures for German and English editorial articles

topics are covered. These differences make it impossible to compare topic models trained on different datasets.

# 5 Future Work and Conclusion

## 5.1 Future work

As this thesis covered multiple areas in the field of topic modeling, the recommendations for possible future improvements are also divided into two sections. We have applied the intrinsic automatic topic labeling technique of Mei et al., 2007. The authors propose several possible improvements that would also improve the labeling on our data. Foremost, a different method for generating canidate labels and matching these to topics is necessary. The extrinsic labeling approach relied on WordNet and could thus only be applied on the English data. The methods could be extended to support German data, for example by incorporating GermaNet or other lexical sources. Lastly, it would be interesting to generate labels for hierarchical topic models.

In this thesis the internal consistency of topic models was studied to analyze the quality of the topics and what effect the chosen number of topics has. In future revisions also the effect of changing other parameters such as the priors $\alpha$ and $\beta$ for LDA on the quality of the modeling should be considered.

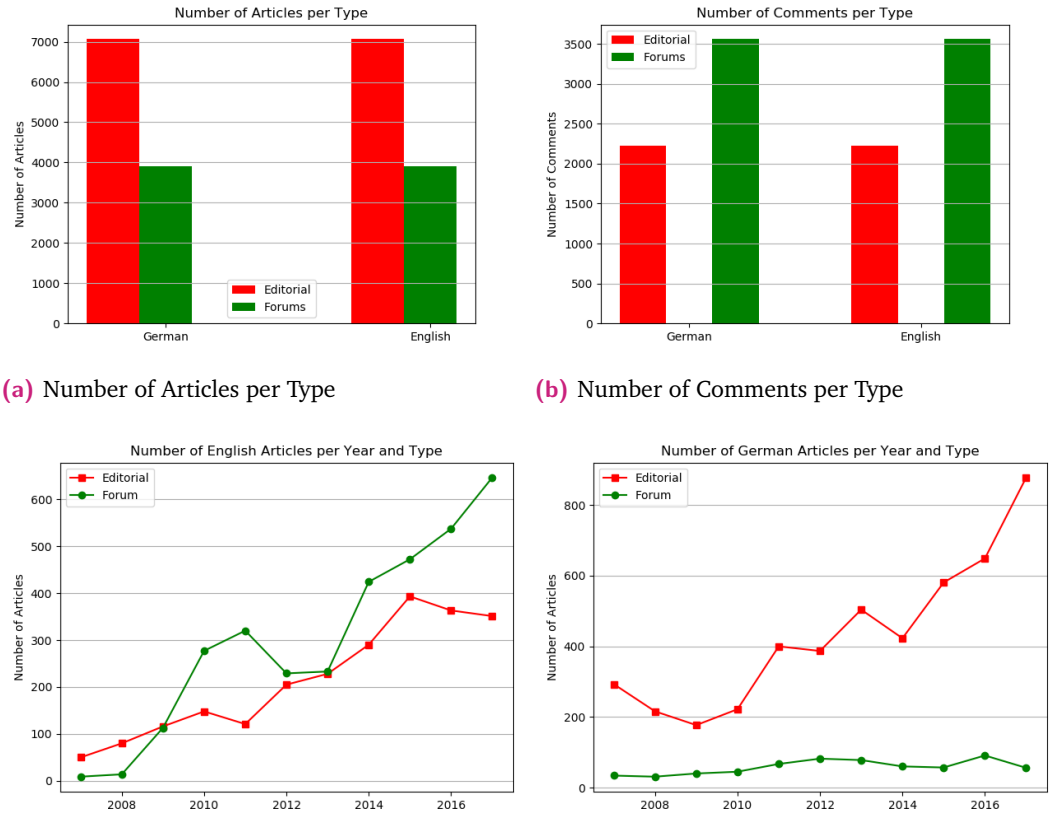
## 5.2 Conclusion

In this thesis two main topics were covered: How can we label topics automatically and how can we measure the internal consistency of topic modeling.

To answer the fist question the intrinsic and extrinsic automatic topic labeling was introduced an evaluated on English and German editorial articles. The evaluation showed, that the intrinsic approach produced meaningful labels on their own, but theses did not fit to the topics. The extrinsic approach was more successful. On average with this method labels were generated, which were fitting more to the topics and some of the automatic labels even matched with the labels, that were submitted by the domain experts.

To answer the second question, several key figures were introduced to measure the internal consistency of topic models with a different number of topic numbers. All key figures were applied on our datasets. It was determined that a single key figure is not enough to specify the number of topics or the quality of the topics. When comparing topic models that were trained on the same dataset, however, the key figures can be used to evaluate which model has more generic or specific topics and how the topics change when varying the number of topics.

# A Descriptive Statistics of the Dataset




(a) Number of Articles per Type

(b) Number of Comments per Type

(c) Distribution of English articles per year

(d) Distribution of German articles per year

**Fig. A.1.:** Descriptive Statistics for all datasets

## A.1 Detailed Statistics of all Sources

## A.2 JSON Storage Schema

[1]The average number of tokens after lemmatizing and stop word removal.

| Source | Total articles | Relevant articles | % rel. articles | Avg. article length[1] | Rel. art. w/ cmnt. | % rel. art. w/ cmnt. |
|---|---|---|---|---|---|---|
| usatoday | 95 | 61 | 64.21 | 303 | 15 | 24.59 |
| nytimes | 438 | 327 | 74.66 | 528 | 99 | 30.28 |
| nypost | 106 | 33 | 31.13 | 377 | 0 | 0.00 |
| washingtonpost | 1563 | 489 | 31.29 | 480 | 285 | 58.28 |
| latimes | 1522 | 270 | 17.74 | 419 | 8 | 2.96 |
| chicagotribune | 2283 | 572 | 25.05 | 420 | 39 | 6.82 |
| huffingtonpost | 880 | 668 | 75.91 | 479 | 0 | 0.00 |
| organicauthority | 66 | 43 | 65.15 | 626 | 0 | 0.00 |

**Tab. A.1.:** Article statistics for English editorial data

| Source | Total comments | Relevant comments | % rel. cmnt. | Root cmnt. | % root cmnt. | Avg. # cmnt. | Avg. cmnt. length[1] |
|---|---|---|---|---|---|---|---|
| usatoday | 259 | 195 | 75.29 | 103 | 52.82 | 3 | 17 |
| nytimes | 16128 | 11576 | 71.78 | 7353 | 63.52 | 35 | 40 |
| nypost | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| washingtonpost | 84669 | 14875 | 17.57 | 6667 | 44.82 | 30 | 24 |
| latimes | 374 | 14 | 3.74 | 12 | 85.71 | 0 | 34 |
| chicagotribune | 281 | 154 | 54.80 | 131 | 85.06 | 0 | 19 |
| huffingtonpost | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| organicauthority | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |

**Tab. A.2.:** Comment statistics for English editorial data

| Source | Total articles | Relevant articles | % rel. articles | Avg. article length[1] | Rel. art. w/ cmnt. | % rel. art. w/ cmnt. |
|---|---|---|---|---|---|---|
| reddit | 256 | 225 | 87.89 | 49 | 190 | 84.44 |
| usmessageboard | 382 | 61 | 15.97 | 0 | 61 | 100.00 |
| cafemom | 88 | 26 | 29.55 | 251 | 26 | 100.00 |
| quora | 1703 | 1497 | 87.90 | 5 | 1304 | 87.11 |
| fb | 5035 | 1467 | 29.14 | 23 | 1355 | 92.37 |

**Tab. A.3.:** Article statistics for English forum data

| Source | Total comments | Relevant comments | % rel. cmnt. | Root cmnt. | % root cmnt. | Avg. # cmnt. | Avg. cmnt. length[1] |
|---|---|---|---|---|---|---|---|
| reddit | 9291 | 8392 | 90.32 | 1574 | 18.76 | 37 | 25 |
| usmessageboard | 78303 | 1982 | 2.53 | 1254 | 63.27 | 32 | 43 |
| cafemom | 2206 | 352 | 15.96 | 280 | 79.55 | 13 | 30 |
| quora | 9606 | 8699 | 90.56 | 5229 | 60.11 | 5 | 46 |
| fb | 299126 | 81660 | 27.30 | 64183 | 78.60 | 55 | 11 |

**Tab. A.4.:** Comment statistics for English forum data

| Source | Total articles | Relevant articles | % rel. articles | Avg. article length[1] | Rel. art. w/ cmnt. | % rel. art. w/ cmnt. |
|---|---|---|---|---|---|---|
| spiegel | 468 | 152 | 32.48 | 376 | 61 | 40.13 |
| zeit | 154 | 62 | 40.26 | 461 | 35 | 56.45 |
| welt | 729 | 392 | 53.77 | 323 | 35 | 8.93 |
| taz | 2458 | 1406 | 57.20 | 255 | 249 | 17.71 |
| tagesspiegel | 625 | 278 | 44.48 | 279 | 41 | 14.75 |
| handelsblatt | 567 | 286 | 50.44 | 302 | 65 | 22.73 |
| freitag | 16 | 7 | 43.75 | 678 | 5 | 71.43 |
| tagesschau | 61 | 17 | 27.87 | 202 | 17 | 100.00 |
| br | 191 | 93 | 48.69 | 297 | 26 | 27.96 |
| wdr | 68 | 37 | 54.41 | 241 | 0 | 0.00 |
| swr | 164 | 82 | 50.00 | 207 | 0 | 0.00 |
| ndr | 18 | 5 | 27.78 | 209 | 0 | 0.00 |
| derstandard | 1092 | 646 | 59.16 | 231 | 529 | 81.89 |
| diepresse | 304 | 152 | 50.00 | 230 | 100 | 65.79 |
| kurier | 287 | 165 | 57.49 | 199 | 88 | 53.33 |
| nachrichtenat | 254 | 134 | 52.76 | 198 | 75 | 55.97 |
| salzburgcom | 154 | 93 | 60.39 | 177 | 0 | 0.00 |
| krone | 97 | 31 | 31.96 | 143 | 0 | 0.00 |
| tagesanzeiger | 187 | 32 | 17.11 | 171 | 17 | 53.12 |
| nzz | 316 | 108 | 34.18 | 338 | 17 | 15.74 |
| aargauer | 110 | 46 | 41.82 | 221 | 17 | 36.96 |
| luzernzeitung | 105 | 55 | 52.38 | 217 | 0 | 0.00 |
| srf | 147 | 85 | 57.82 | 194 | 56 | 65.88 |
| forum_ernaehrung | 18 | 3 | 16.67 | 339 | 0 | 0.00 |
| heise | 33 | 17 | 51.52 | 479 | 17 | 100.00 |
| eatsmarter | 300 | 100 | 33.33 | 176 | 35 | 35.00 |
| huffingtonpost_de | 293 | 94 | 32.08 | 248 | 0 | 0.00 |
| waz | 744 | 207 | 27.82 | 193 | 68 | 32.85 |
| merkur | 393 | 243 | 61.83 | 209 | 69 | 28.40 |
| rp | 604 | 267 | 44.21 | 204 | 103 | 38.58 |
| focus | 777 | 397 | 51.09 | 176 | 154 | 38.79 |
| campact | 61 | 23 | 37.70 | 224 | 23 | 100.00 |

**Tab. A.5.:** Article statistics for German editorial data

| Source | Total comments | Relevant comments | % rel. cmnt. | Root cmnt. | % root cmnt. | Avg. # cmnt. | Avg. cmnt. length[1] |
|---|---|---|---|---|---|---|---|
| spiegel | 62860 | 21551 | 34.28 | 5863 | 27.21 | 141 | 48 |
| zeit | 8496 | 2977 | 35.04 | 1279 | 42.96 | 48 | 32 |
| welt | 1448 | 528 | 36.46 | 316 | 59.85 | 1 | 21 |
| taz | 5537 | 2608 | 47.10 | 1310 | 50.23 | 1 | 28 |
| tagesspiegel | 3535 | 1279 | 36.18 | 1279 | 100.00 | 4 | 36 |
| handelsblatt | 923 | 295 | 31.96 | 222 | 75.25 | 1 | 28 |
| freitag | 129 | 65 | 50.39 | 33 | 50.77 | 9 | 34 |
| tagesschau | 4377 | 841 | 19.21 | 841 | 100.00 | 49 | 32 |
| br | 386 | 343 | 88.86 | 220 | 64.14 | 3 | 26 |
| wdr | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| swr | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| ndr | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| derstandard | 80715 | 50790 | 62.93 | 12152 | 23.93 | 78 | 15 |
| diepresse | 3015 | 1796 | 59.57 | 891 | 49.61 | 11 | 22 |
| kurier | 870 | 471 | 54.14 | 308 | 65.39 | 2 | 17 |
| nachrichtenat | 1992 | 678 | 34.04 | 310 | 45.72 | 5 | 14 |
| salzburgcom | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| krone | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| tagesanzeiger | 4872 | 1139 | 23.38 | 664 | 58.30 | 35 | 18 |
| nzz | 622 | 162 | 26.05 | 101 | 62.35 | 1 | 32 |
| aargauer | 397 | 262 | 65.99 | 122 | 46.56 | 5 | 18 |
| luzernzeitung | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| srf | 1477 | 941 | 63.71 | 652 | 69.29 | 11 | 20 |
| forum_ernaehrung | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| heise | 3636 | 1835 | 50.47 | 335 | 18.26 | 107 | 53 |
| eatsmarter | 1179 | 162 | 13.74 | 146 | 90.12 | 1 | 30 |
| huffingtonpost_de | 0 | 0 | 0.00 | 0 | 0.00 | 0 | 0 |
| waz | 1827 | 459 | 25.12 | 327 | 71.24 | 2 | 25 |
| merkur | 699 | 347 | 49.64 | 194 | 55.91 | 1 | 15 |
| rp | 1808 | 822 | 45.46 | 822 | 100.00 | 3 | 35 |
| focus | 5806 | 2477 | 42.66 | 2123 | 85.71 | 6 | 24 |
| campact | 2577 | 687 | 26.66 | 518 | 75.40 | 29 | 30 |

**Tab. A.6.:** Comment statistics for German editorial data

| Source | Total articles | Relevant articles | % rel. articles | Avg. article length[1] | Rel. art. w/ cmnt. | % rel. art. w/ cmnt. |
|---|---|---|---|---|---|---|
| reddit_de | 83 | 44 | 53.01 | 3 | 33 | 75.00 |
| gutefrage | 547 | 396 | 72.39 | 17 | 396 | 100.00 |
| werweisswas | 33 | 27 | 81.82 | 30 | 26 | 96.30 |
| glamour | 3 | 2 | 66.67 | 58 | 2 | 100.00 |
| webkoch | 4 | 3 | 75.00 | 221 | 2 | 66.67 |
| chefkoch | 248 | 150 | 60.48 | 54 | 150 | 100.00 |
| paradisi | 18 | 18 | 100.00 | 19 | 18 | 100.00 |
| kleiderkreisel | 69 | 24 | 34.78 | 50 | 24 | 100.00 |
| biooekoforum | 1 | 1 | 100.00 | 19 | 1 | 100.00 |
| bfriendsBrigitte | 20 | 11 | 55.00 | 56 | 11 | 100.00 |
| schule-und-familie | 2 | 2 | 100.00 | 32 | 1 | 50.00 |

**Tab. A.7.:** Article statistics for German forum data

| Source | Total comments | Relevant comments | % rel. cmnt. | Root cmnt. | % root cmnt. | Avg. # cmnt. | Avg. cmnt. length[1] |
|---|---|---|---|---|---|---|---|
| reddit_de | 1665 | 488 | 29.31 | 138 | 28.28 | 11 | 16 |
| gutefrage | 6005 | 4100 | 68.28 | 1898 | 46.29 | 10 | 19 |
| werweisswas | 241 | 195 | 80.91 | 195 | 100.00 | 7 | 39 |
| glamour | 287 | 188 | 65.51 | 188 | 100.00 | 94 | 29 |
| webkoch | 34 | 34 | 100.00 | 34 | 100.00 | 11 | 22 |
| chefkoch | 9804 | 5750 | 58.65 | 5750 | 100.00 | 38 | 36 |
| paradisi | 63 | 63 | 100.00 | 63 | 100.00 | 3 | 17 |
| kleiderkreisel | 4831 | 1255 | 25.98 | 854 | 68.05 | 52 | 18 |
| biooekoforum | 15 | 15 | 100.00 | 15 | 100.00 | 15 | 23 |
| bfriendsBrigitte | 2898 | 740 | 25.53 | 740 | 100.00 | 67 | 37 |
| schule-und-familie | 28 | 28 | 100.00 | 28 | 100.00 | 14 | 31 |

**Tab. A.8.:** Comment statistics for German forum data

```
{
   "article_title":"article title",
   "article_author":[
   {
      "article_author_id":"123456789",
      "article_author_name":"author name"
   }
   ],
   "article_time":"2015-10-17 20:02:54",
   "article_text":"article text",
   "article_source":"news source",
   "comments":[
   {
      "comment_id":"123456789",
      "comment_author": {
         "comment_author_id":"45678",
         "comment_author_name":"author name",
      },
      "comment_time":"2015-10-20 04:17:17",
      "comment_text":"comment text",
      "comment_rating":-15.0,
      "comment_title":"example title"
   },
   {
      "comment_id":"987654321",
      "comment_author":{
         "comment_author_id":"12345",
         "comment_author_name":"author name"
      },
      "comment_time":"2015-10-19 19:16:33",
      "comment_text":"comment text",
      "comment_replyTo":"123456789",
      "comment_rating":6.0
   }
   ],
   "search_query":"organic farming",
   "article_url":"https://example.url",
   "resource_type":"editorial | blog | forum",
   "article_rating":5.0
}
```

**Listing 1:** JSON Storage Schema

# Statistics of Internal Consistency

# B

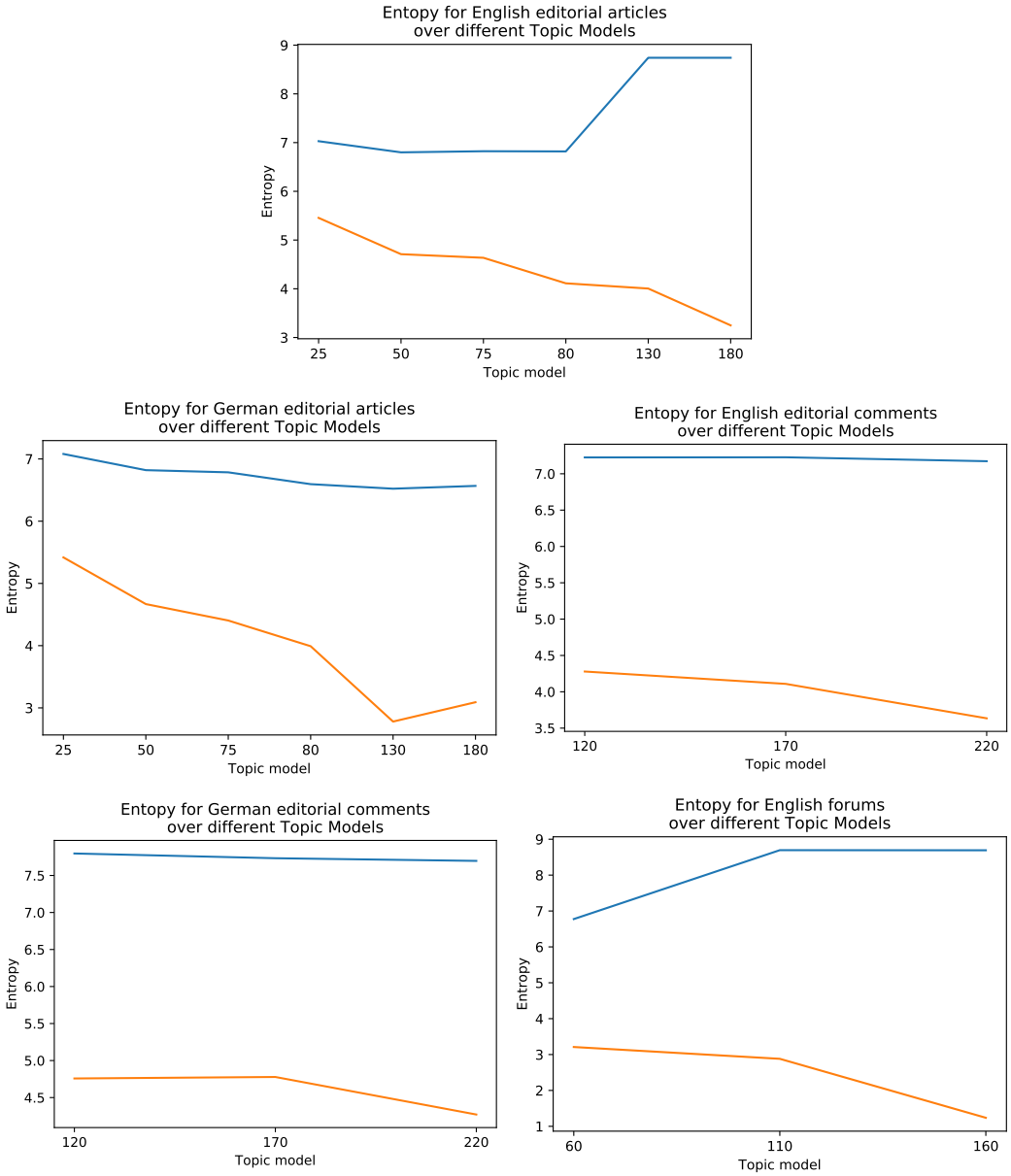



**Fig. B.1.:** Minimal and maximal entropy values for different topic models

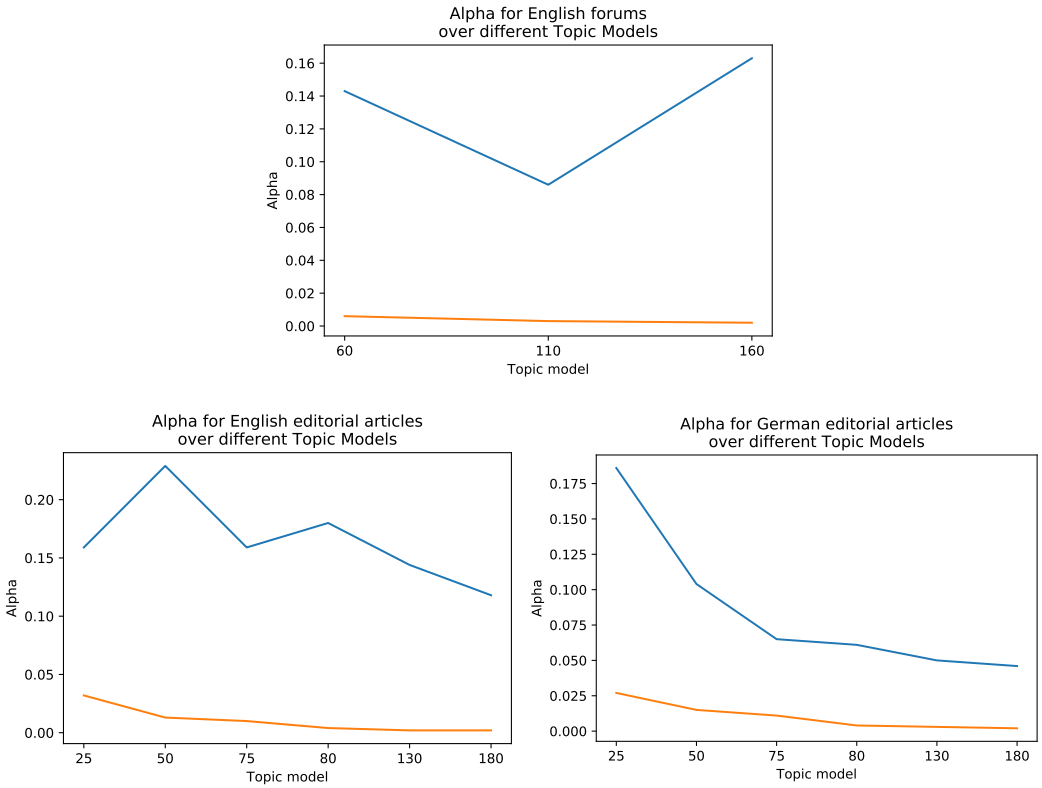



**Fig. B.2.:** Minimal and maximal alpha values for different topic models

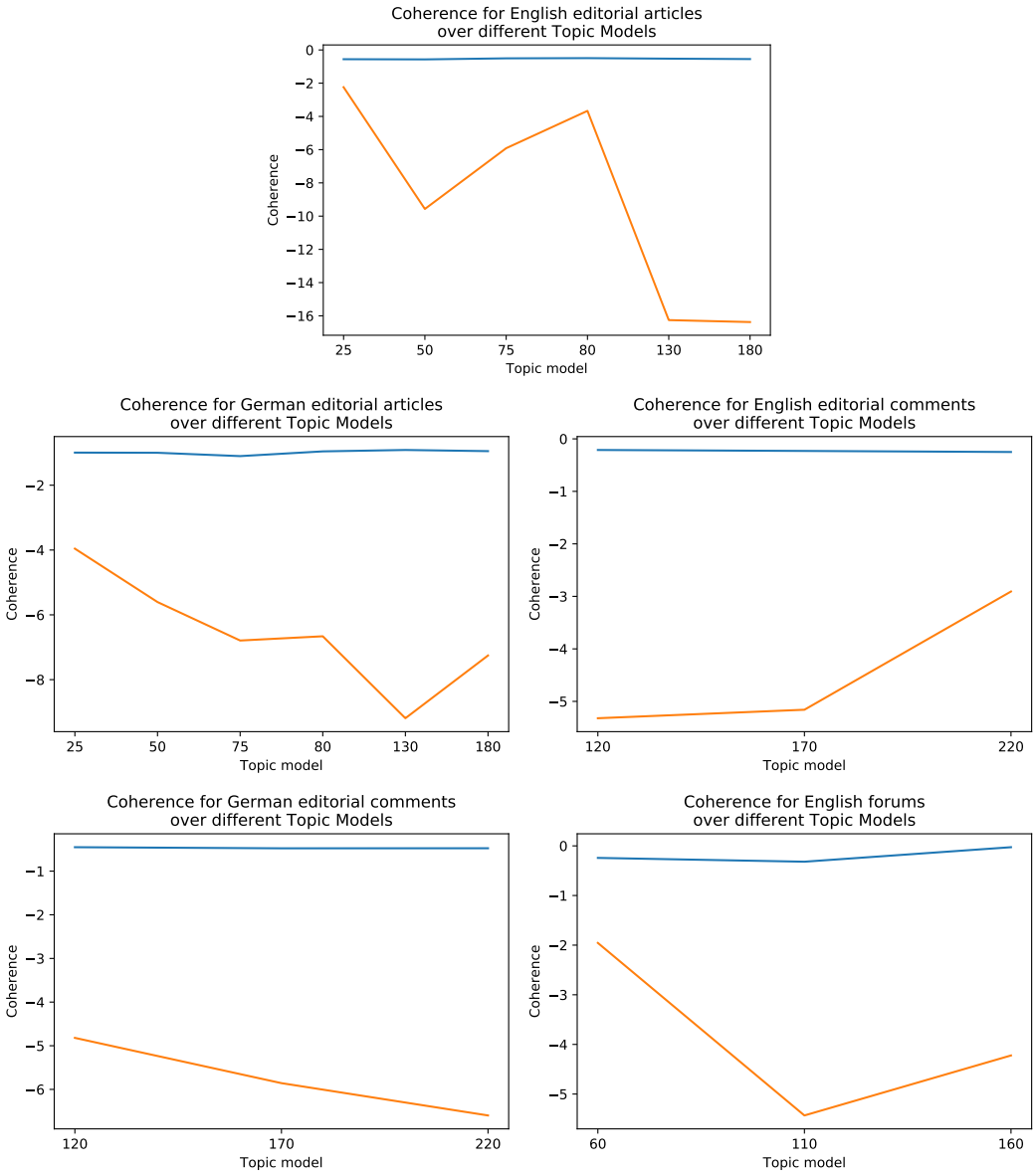



**Fig. B.3.:** Minimal and maximal coherence for different topic models

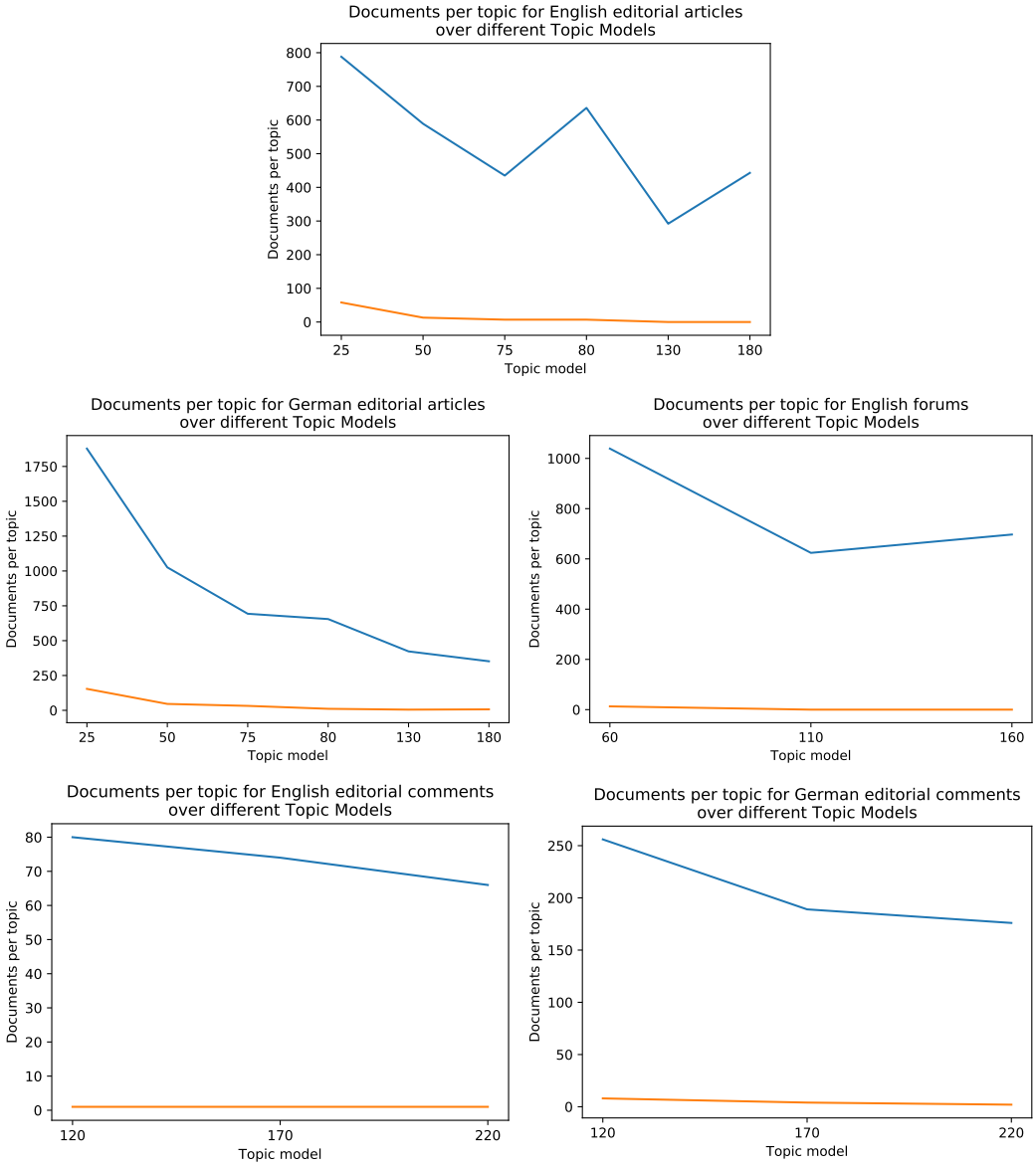



**Fig. B.4.:** Minimal and maximal documents per topic for different topic models

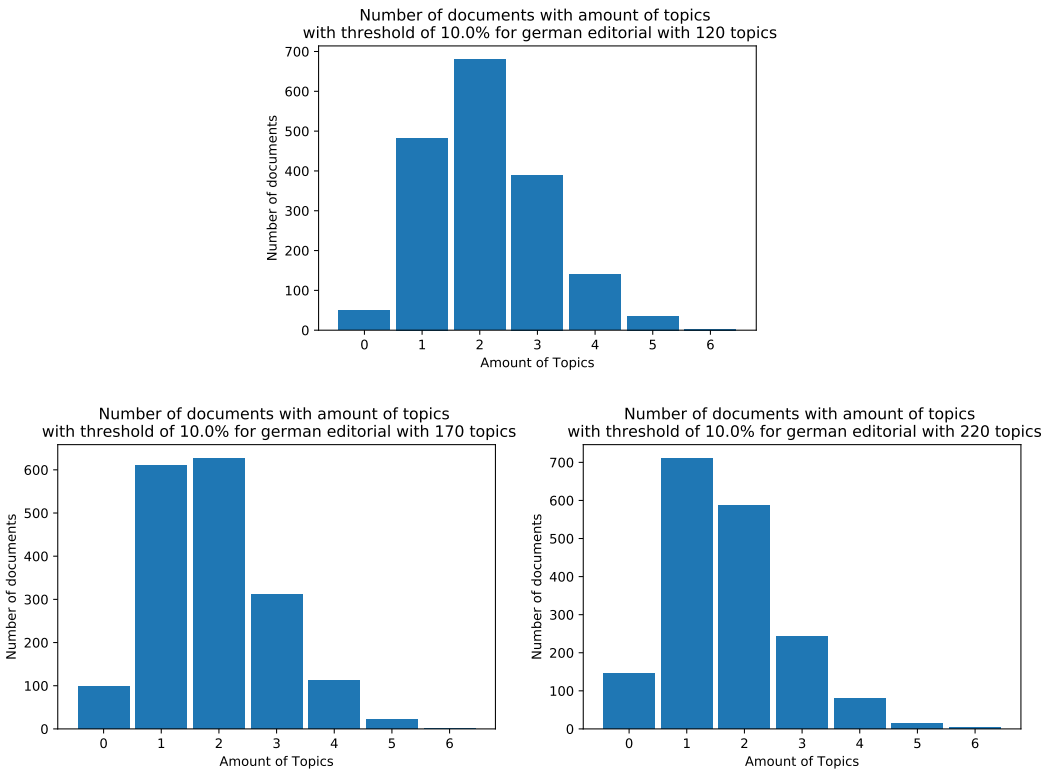



**Fig. B.5.:** Amount of topics per documents for German editorial comments

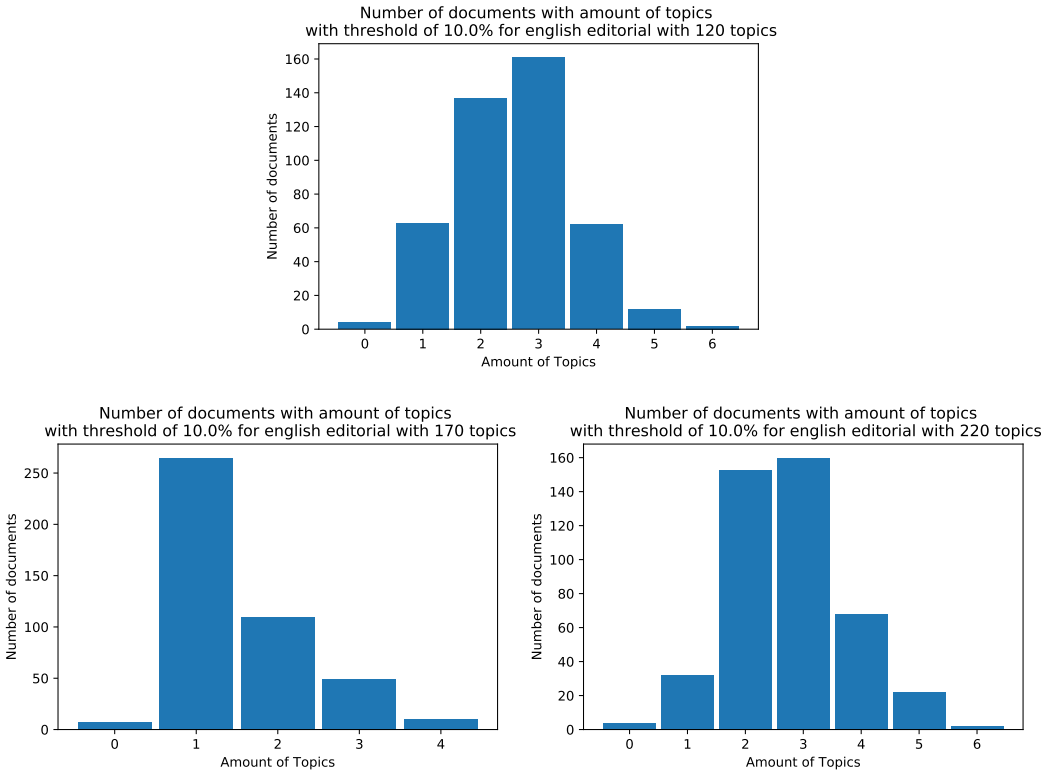



**Fig. B.6.:** Amount of topics per documents for English editorial comments

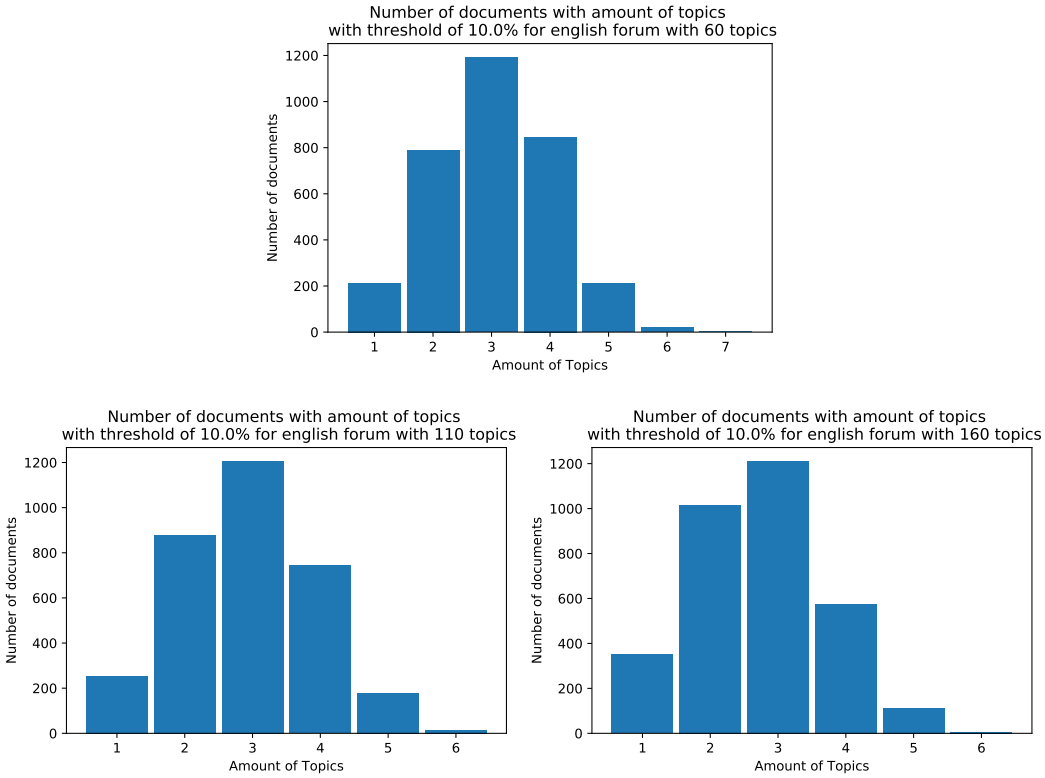



**Fig. B.7.:** Amount of topics per documents for English forums

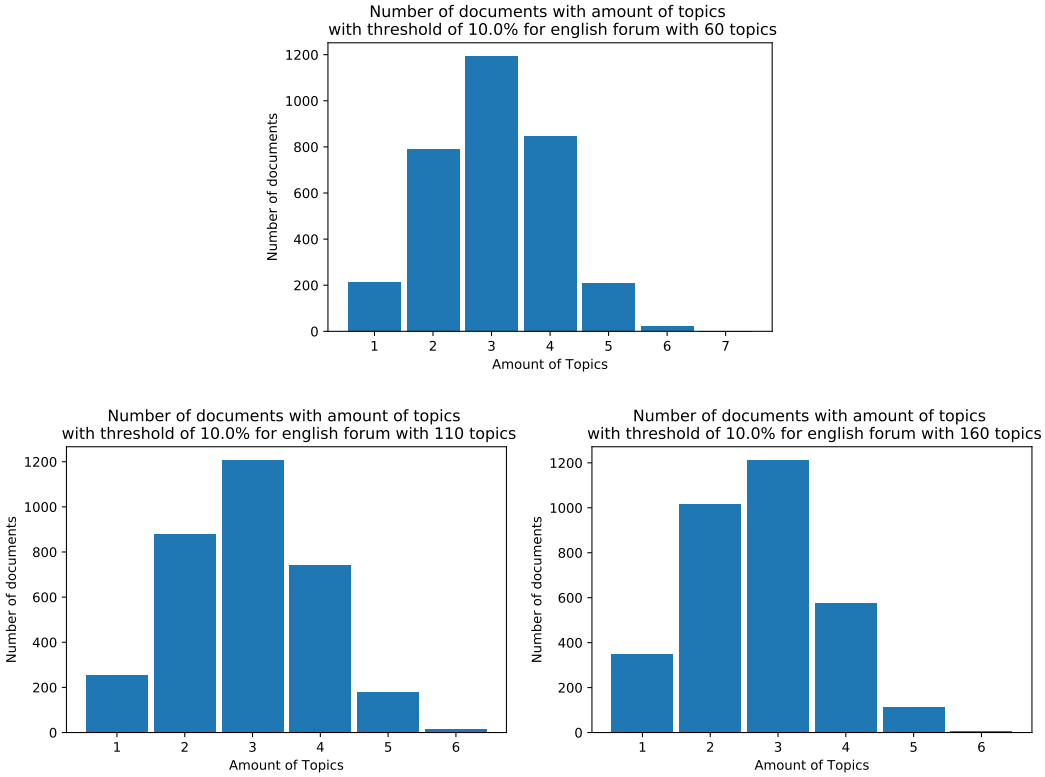
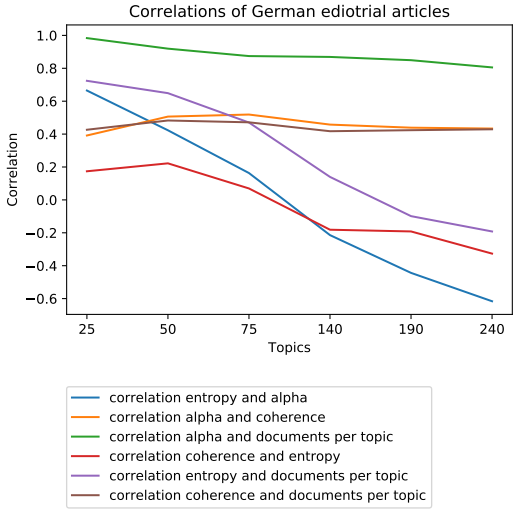
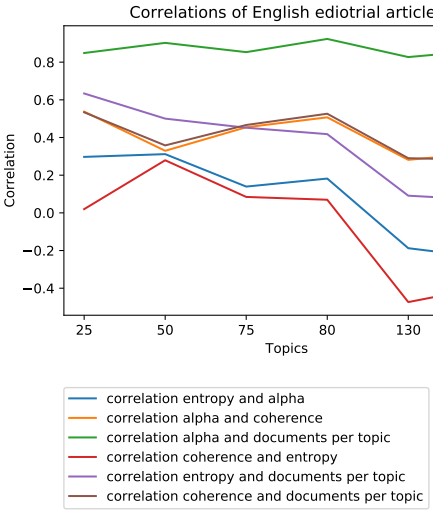


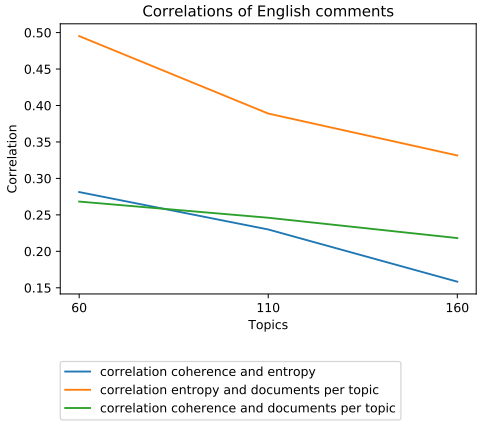

**Fig. B.8.:** Amount of topics per documents for German forums



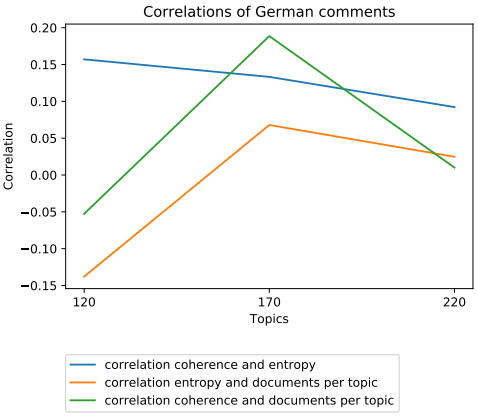

(a) Correlation for German editorial articles



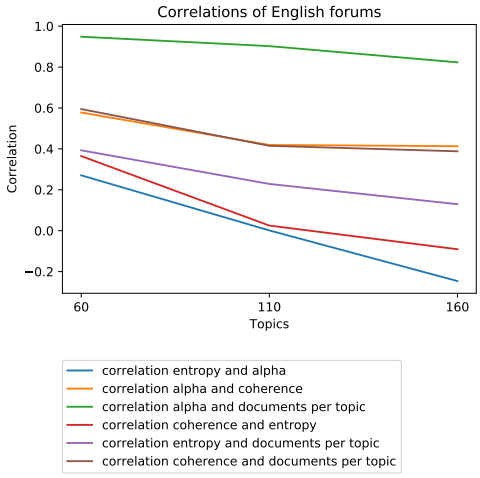

(b) Correlation for English editorial articles




(c) Correlation for English comments

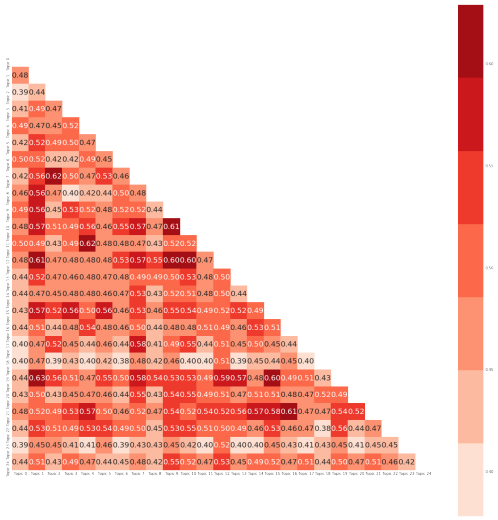



(d) Correlation for German comments

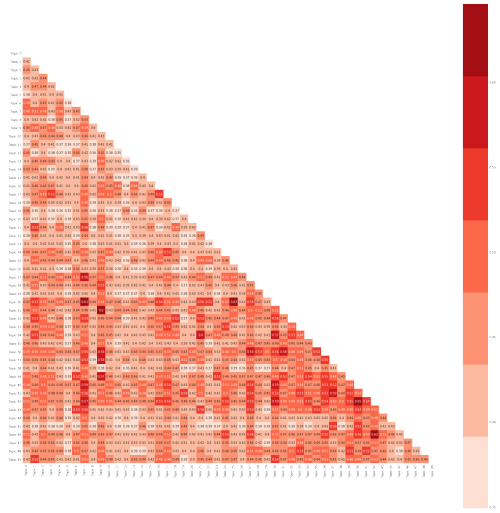



(e) Correlation for English forums

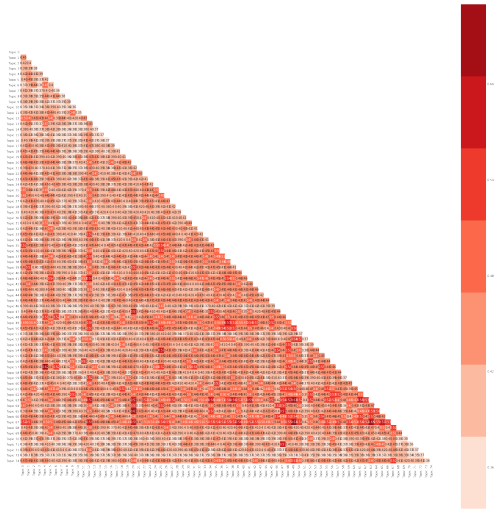
**Fig. B.9.:** Correlations for different topic models



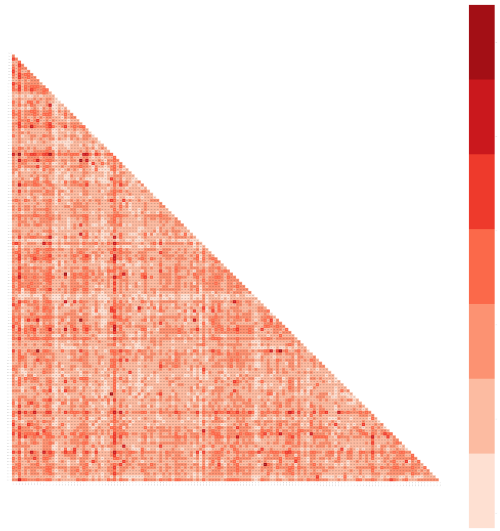
(a) Heat map for the topic model with 25 topics



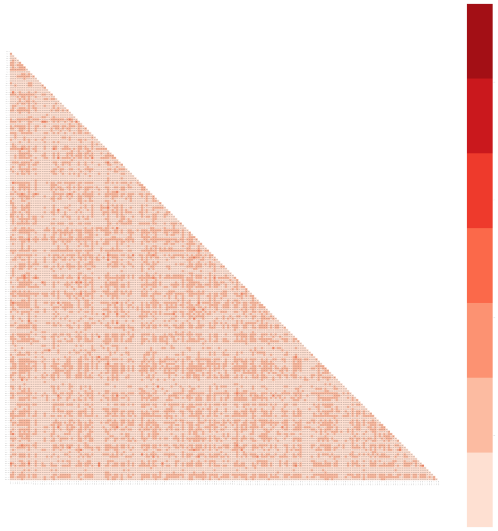
(b) Heat map for the topic model with 50 topics



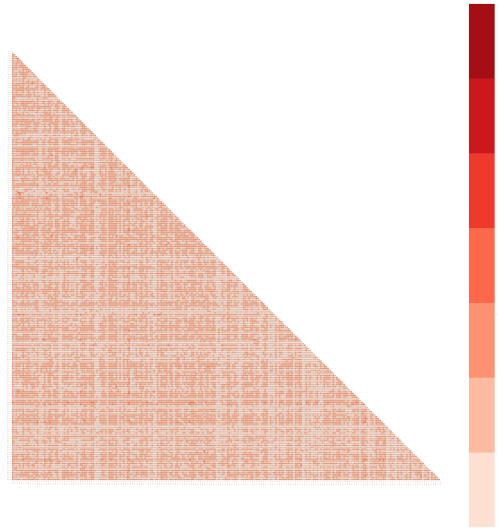
(c) Heat map for the topic model with 75 topics



(d) Heat map for the topic model with 140 topics

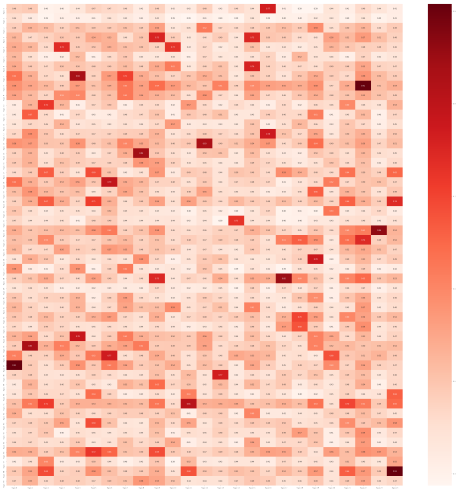


(e) Heat map for the topic model with 190 topics

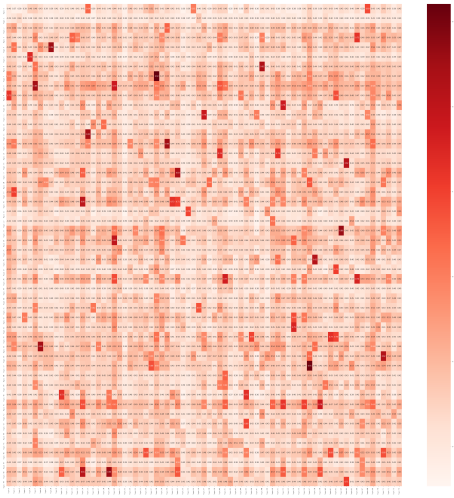


(f) Heat map for the topic model with 240 topics

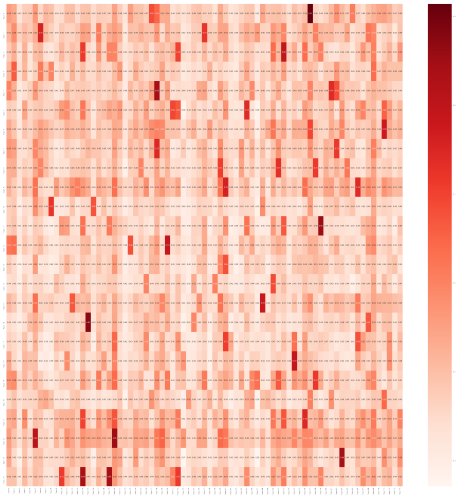
**Fig. B.10.:** Heat maps with similarities intra a topic models for German editorial articles



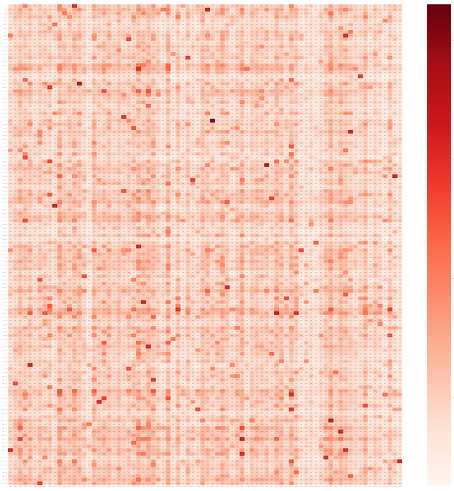
(a) Heat map for topic models with 25 and 50 topics



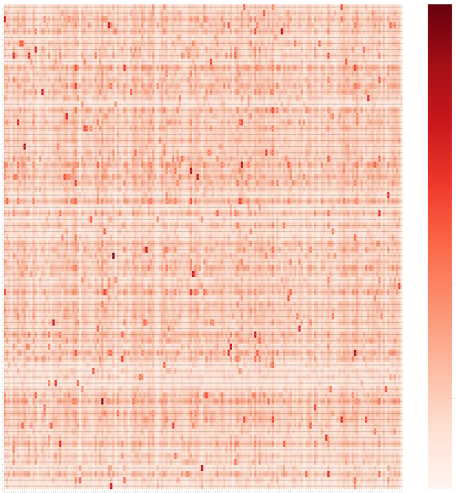
(b) Heat map for topic models with 50 and 75 topics



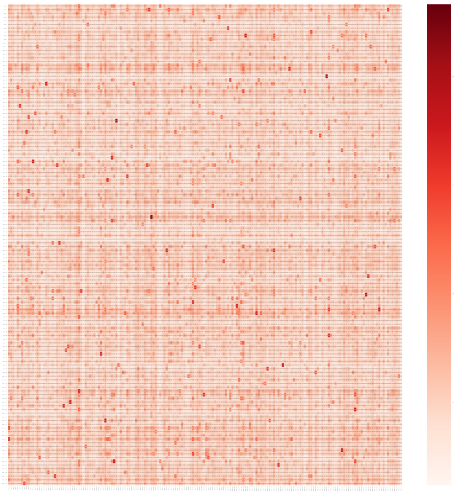
(c) Heat map for topic models with 25 and 75 topics



(d) Heat map for topic models with 80 and 130 topics

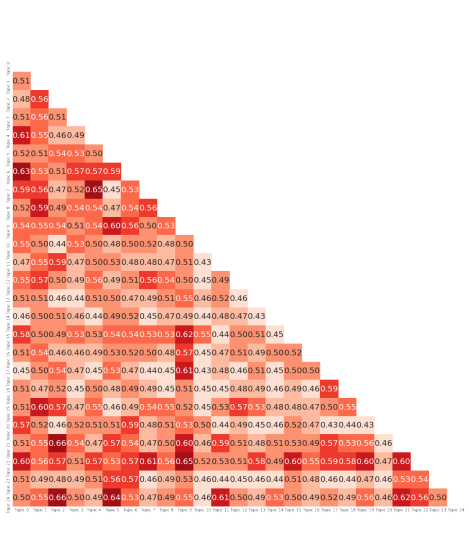


(e) Heat map for topic models with 80 and 180 topics

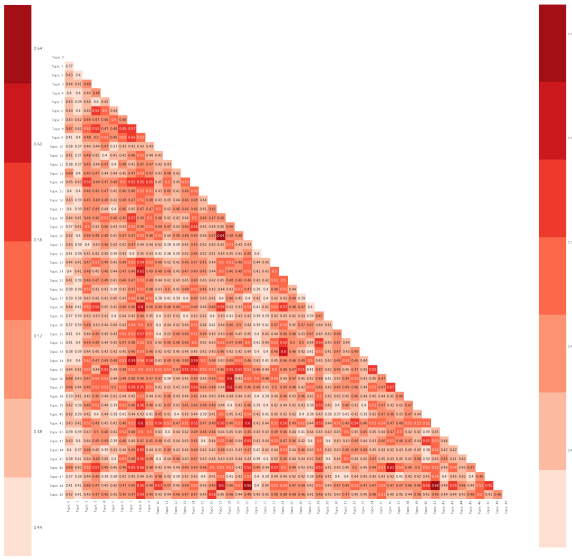


(f) Heat map for the topic models with 130 and 180 topics

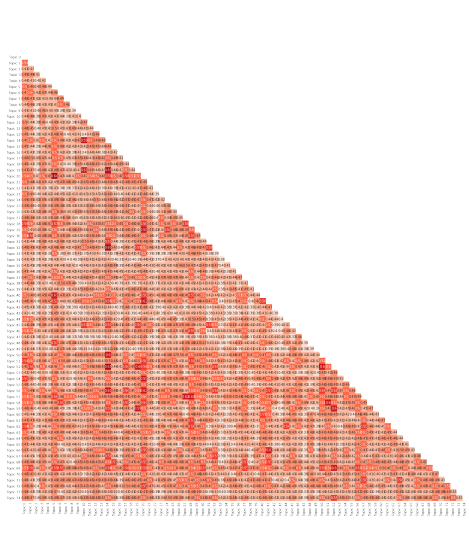
**Fig. B.11.:** Heat maps with similarities inter two topic models for English editorial articles



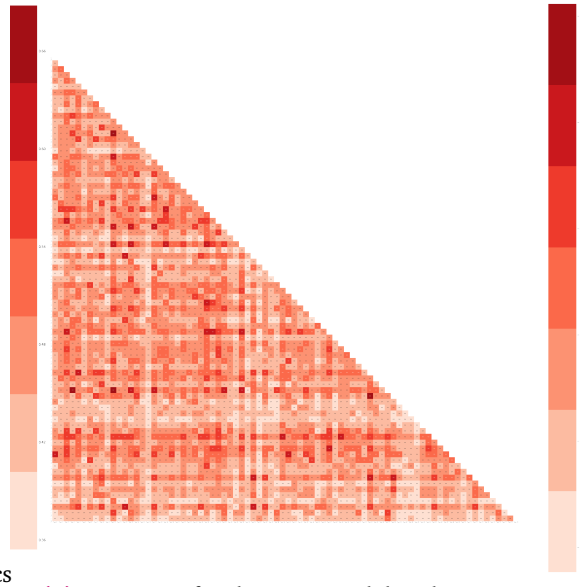
(a) Heat map for the topic model with 25 topics



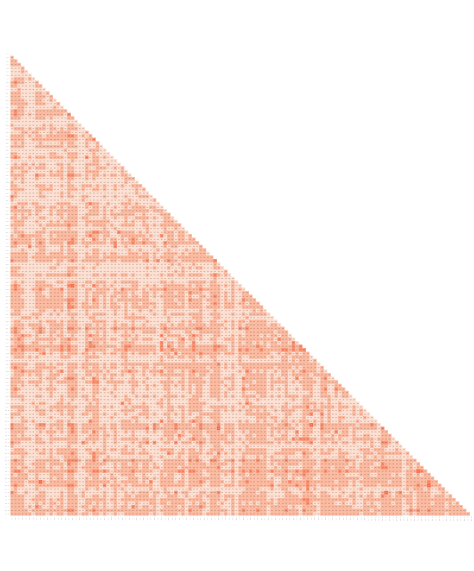
(b) Heat map for the topic model with 50 topics



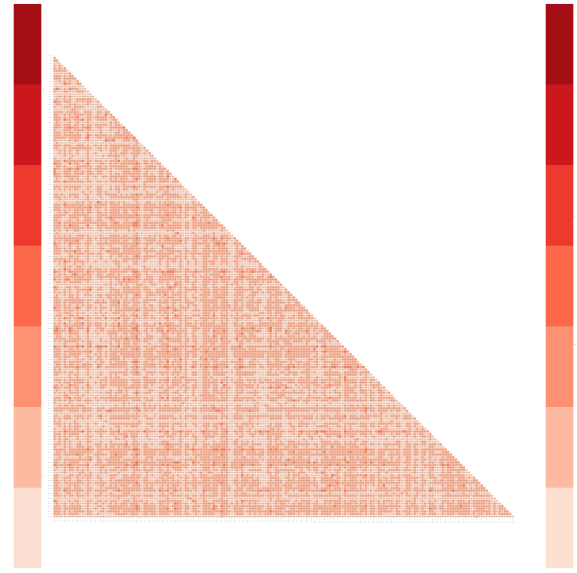
(c) Heat map for the topic model with 75 topics



(d) Heat map for the topic model with 80 topics

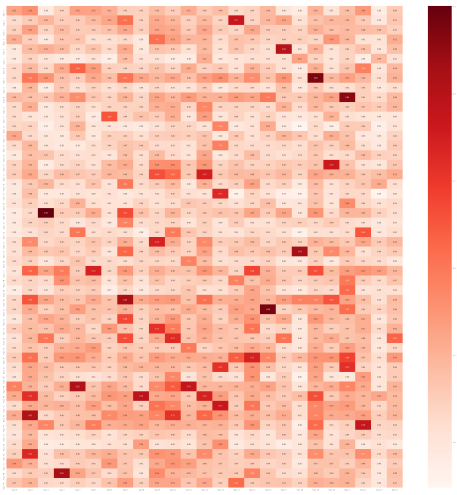


(e) Heat map for the topic model with 130 topics

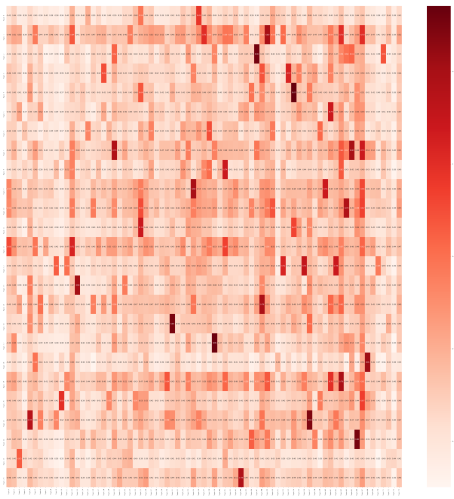


(f) Heat map for the topic model with 180 topics

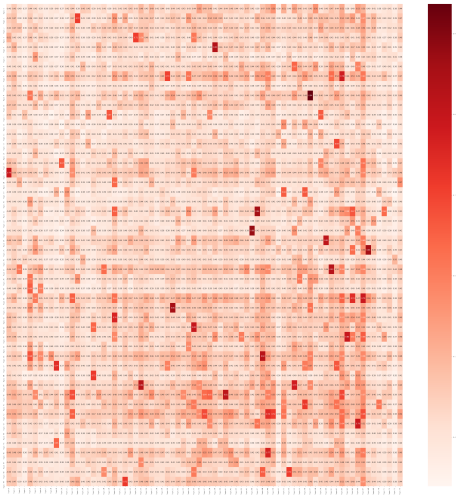
**Fig. B.12.:** Heat maps with similarities intra two topic models for English editorial articles



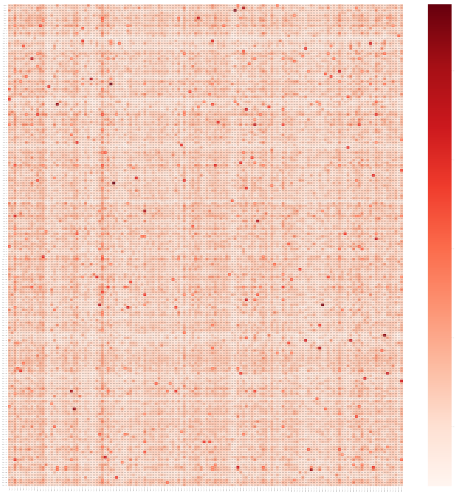
(a) Heat map for topic models with 25 and 50 topics



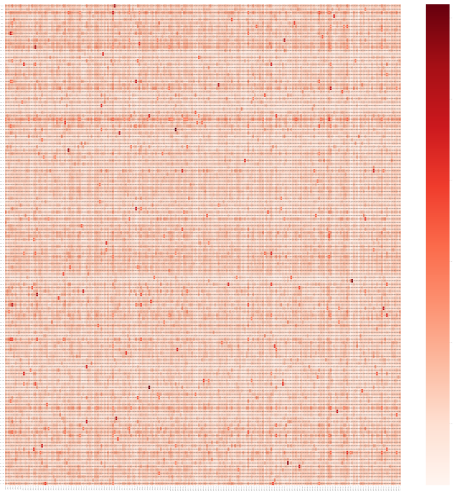
(b) Heat map for topic models with 25 and 75 topics



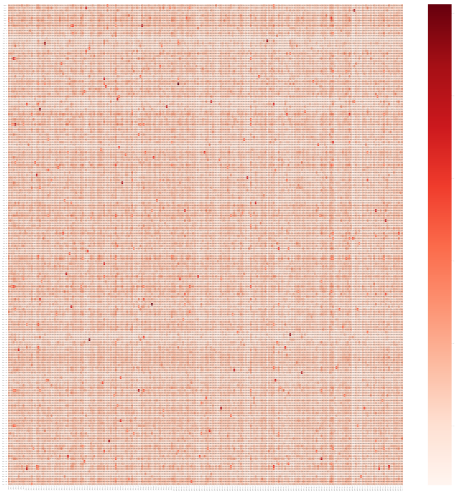
(c) Heat map for topic models with 50 and 75 topics



(d) Heat map for topic models with 140 and 190 topics



(e) Heat map for topic models with 140 and 240 topics



(f) Heat map for the topic models with 190 and 240 topics

**Fig. B.13.:** Heat maps with similarities inter two topic models for German editorial articles

# Bibliography


AGOF (2018). *Nettoreichweite der Top 15 Nachrichtenseiten (ab 14 Jahre) im November 2014 in Unique Usern (in Millionen)* (cit. on p. 12).

Allahyari, Mehdi and Krys Kochut (2015). „Automatic Topic Labeling using Ontology-based Topic Models“. In: (cit. on p. 17).

Ankit Sethi, Bharat Upadrasta (2012). „Introduction to Probabilistic Topic Modeling“. In: 151, pp. 10–17 (cit. on p. 34).

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin (2016). „Automatic Labelling of Topics with Neural Embeddings“. In: 1, pp. 953–963. arXiv: 1612.05340 (cit. on pp. 18, 21).

Blei (2003). „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research* 3.3/1/2003, pp. 993–1022. arXiv: 1111.6189v1 (cit. on pp. 7, 9, 33).

Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov (2004). „Metagenes and molecular pattern discovery using matrix factorization“. In: *Proceedings of the National Academy of Sciences* (cit. on p. 10).

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990). „Indexing by latent semantic analysis“. In: *Journal of the American Society for Information Science*. arXiv: arXiv:1403.2923v1 (cit. on p. 7).

Fei-Fei, Li and Pietro Perona (2005). „A bayesian hierarchical model for learning natural scene categories“. In: *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*. arXiv: 0305217 [math.AP] (cit. on p. 8).

Griffiths, T. L. and M. Steyvers (2004). „Finding scientific topics“. In: *Proceedings of the National Academy of Sciences* (cit. on p. 3).

Griffiths, Thomas L. and Mark Steyvers (2002). „A probabilistic approach to semantic representation“. In: *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (cit. on p. 9).

Hofmann, Thomas (2001). „Unsupervised learning by probabilistic Latent Semantic Analysis“. In: *Machine Learning*. arXiv: 0005074v1 [arXiv:astro-ph] (cit. on p. 7).

Hulpus, Ioana, Conor Hayes, Marcel Karnstedt, and Derek Greene (2013). „Unsupervised graph-based topic labelling using dbpedia“. In: *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, p. 465 (cit. on pp. 18, 21).

IVW (2018). *Verkaufte Auflage der überregionalen Tageszeitungen in Deutschland im 3 . Quartal 2018* (cit. on p. 12).

Jelodar, Hamed, Yongli Wang, Chi Yuan, and Xia Feng (2017). „Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey“ (cit. on p. 8).

Jurafsky, Daniel and James H Martin (2009). „Speech and Language Processing“. In: *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition* 21, pp. 0–934. arXiv: arXiv:1011.1669v3 (cit. on p. 6).

Kou, Wanqiu, Fang Li, and Timothy Baldwin (2015). „Automatic labelling of topic models using word vectors and letter trigram vectors“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9460.1, pp. 253–264 (cit. on p. 18).

Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin (2011). „Automatic Labelling of Topic Models“. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 1536–1545 (cit. on pp. 17, 18, 21).

Lee, Daniel D. and H. Sebastian Seung (1999). „Learning the parts of objects by non-negative matrix factorization“. In: *Nature*. arXiv: arXiv:1408.1149 (cit. on p. 9).

Li, Li Jia, Chong Wang, Yongwhan Lim, David M. Blei, and Li Fei-Fei (2010). „Building and using a Semantivisual image hierarchy“. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (cit. on p. 8).

Lin, Jianhua (1991). „Divergence Measures Based on the Shannon Entropy“. In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151 (cit. on p. 36).

Magatti, Davide, Silvia Calegari, Davide Ciucci, and Fabio Stella (2009). „Automatic labeling of topics“. In: *ISDA 2009 - 9th International Conference on Intelligent Systems Design and Applications*, pp. 1227–1232 (cit. on pp. 18, 21).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to Information Retrieval*. arXiv: 052186571997805218657 15 (cit. on p. 5).

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). „Automatic labeling of multinomial topic models“. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07* January 2007, p. 490 (cit. on pp. 16, 18–20, 25, 26, 48).

Miller, George A. (1995). „WordNet: a lexical database for English“. In: *Communications of the ACM* 38.11, pp. 39–41 (cit. on p. 22).

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011). „Optimizing Semantic Coherence in Topic Models“. In: *Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11* 2, pp. 262–272 (cit. on p. 35).

Minka, T and J Lafferty (2002). „Expectation-propagation for the generative aspect model“. In: *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence* (cit. on p. 9).

Newman, David, Jh Lau, Karl Grieser, and Timothy Baldwin (2010). „Automatic evaluation of topic coherence“. In: ... *Language Technologies: The* ... June, pp. 100–108 (cit. on p. 35).

Nikolenko, Sergey I., Sergei Koltcov, and Olessia Koltsova (2017). „Topic modelling for qualitative studies“. In: *Journal of Information Science*. arXiv: arXiv:1011.1669v3 (cit. on p. 2).

Pauca, V. Paul, J. Piper, and Robert J. Plemmons (2006). „Nonnegative matrix factorization for spectral data analysis“. In: *Linear Algebra and Its Applications* (cit. on p. 10).

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly (2000). „Inference of population structure using multilocus genotype data“. In: *Genetics*. arXiv: 0208024 [gr-qc] (cit. on p. 8).

Salton, G, A Wong, and C S Yang (1975). „1975.A vector space model for automatic indexing.pdf“. In: 18.11 (cit. on p. 6).

Stevens, Keith;Kegelmeyer,Philip;Andrzejewski, David;Buttler, David (2012). „Exploring Topic Coherence over many models and many topics“. In: (cit. on pp. 35, 36).

Steyvers, M and T Griffiths (2007a). „Probabilistic topic models“. In: *Handbook of Latent Semantic Analysis: A Road to Meaning*, pp. 424–440. arXiv: 1111.6189v1 (cit. on pp. 33, 37).

Steyvers, Mark and Tom Griffiths (2007b). „Probabilistic topic models“. In: *Handbook of latent semantic analysis*. Vol. 427. 7, pp. 424–440. arXiv: 1111.6189v1 (cit. on p. 11).

Widmer, Christian (2018). „Topic Modeling for Opinion Mining“. In: (cit. on pp. 3, 8, 10).

Zhao, Wayne Xin, Jing Jiang, Jing He, et al. (2011). „Topical keyphrase extraction from Twitter“. In: *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 379–388 (cit. on p. 17).

# List of Figures

# List of Tables