

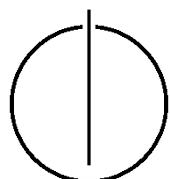
DEPARTMENT OF INFORMATICS

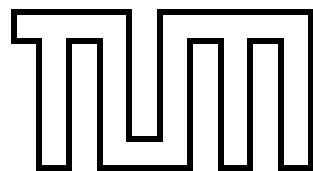
TECHNICAL UNIVERSITY OF MUNICH

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

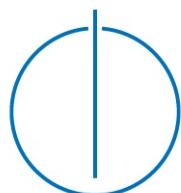
Topic Model Visualisierung für Opinion Mining

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 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 the 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.

# Zusammenfassung

blablablub

# Acknowledgement

As this thesis borders between computer science and qualitative research on consumer behaviour I would like to thank 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

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Thesis structure . . . . .	3
<b>2</b>	<b>Methodology</b>	<b>4</b>
2.1	Document representation . . . . .	4
2.1.1	Bag of Words . . . . .	4
2.1.2	Tf-Idf Weighting . . . . .	4
2.1.3	Vector space model . . . . .	5
2.2	Topic Modeling . . . . .	6
2.2.1	Latent Dirichlet Allocation . . . . .	6
2.2.2	Non negative Matrix Factorization . . . . .	6
2.2.3	Hierarchical Latent Dirichlet Allocation . . . . .	6
<b>3</b>	<b>Dataset</b>	<b>7</b>
3.1	Data collection . . . . .	7
3.2	Data processing . . . . .	8
3.3	Final Datasets . . . . .	8
3.4	Topic Generation . . . . .	9
<b>4</b>	<b>Experiments and Evaluation</b>	<b>11</b>
4.1	Topic ranking . . . . .	11
4.1.1	Related work . . . . .	11
4.1.2	Topic Coherence . . . . .	11
4.1.3	Summed Theta $\theta$ . . . . .	11
4.1.4	Iter-rater reliability . . . . .	11
4.2	Automatic Topic Labeling . . . . .	13
4.2.1	Related work . . . . .	13
4.2.2	Intrinsic Topic Labeling . . . . .	16
4.2.3	Extrinsic Labeling . . . . .	18
4.2.4	Evaluation . . . . .	21
4.3	Internal consistency . . . . .	28
4.3.1	Theta $\theta$ . . . . .	29
4.3.2	Alpha . . . . .	29
4.3.3	Entropy . . . . .	29

4.3.4	Coherence . . . . .	31
4.3.5	Jensen Shannon divergence . . . . .	32
4.3.6	Evaluation . . . . .	33
<b>5</b>	<b>Future Work and Conclusion</b>	<b>45</b>
5.1	Future work . . . . .	45
5.2	Conclusion . . . . .	45
<b>A</b>	<b>Descriptive Statistics of the Dataset</b>	<b>47</b>
A.1	Detailed Statistics of all Sources . . . . .	47
A.2	JSON Storage Schema . . . . .	47
	<b>Bibliography</b>	<b>54</b>

## List of acronyms

<b>ATL</b> Automatic Topic Labeling .....	2
<b>BoW</b> Bag of Words .....	4
<b>Csf</b> Custom scoring function .....	19
<b>HLDA</b> Hierarchical Latent Dirichlet Allocation .....	4
<b>IC</b> Information Content .....	20
<b>IR</b> Information Retrieval .....	5
<b>JS</b> Jensen Shannon .....	32
<b>KL</b> Kullback Leibler .....	15
<b>LDA</b> Latent Dirichlet Allocation .....	2
<b>NLP</b> Natural language processing .....	4
<b>NMF</b> Non-negative Matrix Factorization .....	2
<b>PMI</b> point-wise mutual information .....	14
<b>POS</b> Part-of-speech .....	8
<b>tf-idf</b> term frequency - inverse document frequency .....	5

## 3 Introduction

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

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

31 After ATL another main goal of this work is to prove the internal consistency, which  
32 means, to analyze how the topics itself and the distribution of topics on documents  
33 change when increasing the number of topics. Concretely, it shall be analyzed with  
34 different key figures whether the topics are getting specific or general and what

<sup>35</sup> specific and general in this context means. Furthermore, it shall be examined if the  
<sup>36</sup> topics split up and whether this can be proven according to the top 10 words of  
<sup>37</sup> the topic. All theses analyses shall provide domain experts the overview how topics  
<sup>38</sup> change when increasing or decreasing the number of topics, so they can assess,  
<sup>39</sup> which topic model is the most appropriate one for their expected results.

## <sup>40</sup> 1.1 Thesis structure

<sup>41</sup> First the methods, which were used to identify the topics are introduced in Chapter  
<sup>42</sup> **2**. This includes the approaches to represent the content of documents numerically  
<sup>43</sup> and the algorithms for topic modeling with Latent Dirichlet Allocation (**LDA**) and  
<sup>44</sup> Non-negative Matrix Factorization (**NMF**) .

<sup>45</sup> In the **3**. chapter we introduce the dataset and show how the data were gathered  
<sup>46</sup> and preprocessed.

<sup>47</sup> In the first half of chapter 4, in **4.2**, the possible approaches for Automatic Topic  
<sup>48</sup> Labeling are described and the results of applying those on our dataset is discussed.  
<sup>49</sup> Accordingly, in the second half of chapter 4, in **4.3**, different key figures to measure  
<sup>50</sup> the Internal consistency are first introduced, applied and then discussed on our  
<sup>51</sup> dataset.

<sup>52</sup> Chapter **5** completes the thesis by providing an outlook for possible future work and  
<sup>53</sup> summarizes the thesis with the conclusion.

# 54 Methodology

56 In this chapter the basic principles for the following chapters will be explained.  
57 Section 2.1 describes how documents can be numerically represented and how a  
58 fixed vector representation of documents can be reached. In Section 2.2 then will  
59 explained why Topic models are necessary and three Topic Models LDA, NMF and  
60 Hierarchical Latent Dirichlet Allocation (HLDA) will be describe, which were used in  
61 this thesis.

## 62 2.1 Document representation

### 63 2.1.1 Bag of Words

64 The Bag of Words Bag of Words (BoW) model serves as a numerical representation  
65 of a document, which is used as input for further Natural language processing  
66 (NLP) tasks. It represents the document simply by the counts for each word. The  
67 grammar and the ordering of the words are ignored, so some information is lost.  
68 The document *John likes organic but Mary doesn't* and the document *Mary likes*  
69 *organic but John doesn't* have the same BoW representation although these differ in  
70 context. Nevertheless, similar BoW imply similar document content (Manning et al.,  
71 2008).

### 72 2.1.2 Tf-Idf Weighting

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

$$idf_{d,t} = \log \frac{N_d}{df_{d,t}} \quad (2.1)$$

- <sup>79</sup>  $N_d$  is the number of all documents in a corpus, while  $df_{d,t}$  is the number of documents  
<sup>80</sup> that contain the single term.
- <sup>81</sup> Based on the term frequency  $tf_{t,d}$  and the inverse document frequency  $idf_{d,t}$  we  
<sup>82</sup> introduce the *term frequency - inverse document frequency (tf-idf)*:

$$tf - idf_{d,t} = tf_{t,d} * idf_{d,t} \quad (2.2)$$

- <sup>83</sup> The **tf-idf** weighting has the highest score when the term occurs frequently within a  
<sup>84</sup> small amount of documents. The score is lower when the term occurs rarely or too  
<sup>85</sup> often in many documents (Jurafsky and Martin, 2009).

### <sup>86</sup> 2.1.3 Vector space model

- <sup>87</sup> The representation of documents in the same vector space is known as the vector  
<sup>88</sup> space model. This was originally introduced for Information Retrieval (IR) operations  
<sup>89</sup> like scoring documents on a query, document classification or clustering Salton et al.,  
<sup>90</sup> 1975.

<sup>91</sup> The vector space model forms with the documents  $D_i$  and all unique terms  $T_j$  the  
<sup>92</sup> document term matrix  $C$ . Each row of  $C$  corresponds every single document of the  
<sup>93</sup> corpus and each column the single unique terms. In  $C_{ij}$  the weightings either as  
<sup>94</sup> term frequency or **tf-idf** for each term over all documents is stored.

<sup>95</sup> In Table 2.1 the term frequency and in Table 2.2 **tf-idf** is calculated from three sample  
<sup>96</sup> documents: *Doc 1: Organic is healthier then conventional food*, *Doc 2: I buy organic*  
<sup>97</sup> and *Doc 3: Organic is wasted money*. In this thesis both topic modeling algorithms  
<sup>98</sup> take the document term matrix as input, but with different weightings. For LDA the  
<sup>99</sup> 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.

<sup>100</sup>

	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.

## <sup>101</sup> 2.2 Topic Modeling

<sup>102</sup> Every day large amounts of information are collected and become available. The  
<sup>103</sup> vast quantities of data make it difficult to access those information we are looking  
<sup>104</sup> for. Therefore we need methods that help us to organize, summarize and understand  
<sup>105</sup> large collections of data.

<sup>106</sup> Topic Modeling is used to process large collections efficiently. It helps to discover  
<sup>107</sup> hidden themes or rather topics of document collections. A topic is a multinomial  
<sup>108</sup> distribution over all words in a corpus. Of course the probabilities over each word  
<sup>109</sup> are different.

### <sup>110</sup> 2.2.1 Latent Dirichlet Allocation

### <sup>111</sup> 2.2.2 Non negative Matrix Factorization

### <sup>112</sup> 2.2.3 Hierarchical Latent Dirichlet Allocation

# 114 Dataset

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

## 123 3.1 Data collection

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

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

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

---

<sup>1</sup>only an example German national print press

<sup>2</sup>only an example German news sites

<sup>142</sup> presentations, and stock market information remained. These were kicked out by  
<sup>143</sup> the binary Naive Bayes classifier, which was trained on 1000 random articles<sup>3</sup>, that  
<sup>144</sup> were labeled either as relevant or not by the Domain experts. The final collection  
<sup>145</sup> stored in a JSON schema and the list of all sources and their percentage of relevant  
<sup>146</sup> articles together with other descriptive statistics can be found in Appendix A.

## <sup>147</sup> 3.2 Data processing

<sup>148</sup> For applying further NLP tasks, the extracted dataset was transformed by using  
<sup>149</sup> several pre-processing tasks: First, the texts were tokenized and lowercased. Then  
<sup>150</sup> all common words including numbers and punctuations were removed and Emails  
<sup>151</sup> and Url's were replaced by <EMAIL> and <URL> tags. Second, the remaining  
<sup>152</sup> tokens were lemmatized, so that the inflections of words were replaced by their  
<sup>153</sup> basic form. Third, the texts were examined for collocations, which are co-occurring  
<sup>154</sup> words like *Stiftung Warentest* or *Whole Foods*, with a Gensim library<sup>4</sup>. For the  
<sup>155</sup> lemmatization and tokenization the Spacy library<sup>5</sup> was used. Additionally, in this  
<sup>156</sup> project Part-of-speech (POS)-Tagging was applied to the texts, which is a process  
<sup>157</sup> marking up the words to a particular part of speech, to facilitate the ATL in chapter  
<sup>158</sup> 4.2.

## <sup>159</sup> 3.3 Final Datasets

<sup>160</sup> Before reporting the datasets itself, the definition of text types will be described,  
<sup>161</sup> which were introduced because of the different content and language style. All data  
<sup>162</sup> referring to a main text of a side are called *editorial articles* and the comments under  
<sup>163</sup> the editorial articles are called *editorial comments*. The term *Forum* includes the  
<sup>164</sup> initial question and the comments under it. In this thesis the blogs, which were split  
<sup>165</sup> in editorial and comments, were neglected, because the amount of data and context  
<sup>166</sup> quality was to low.

<sup>167</sup> We created two different final datasets where the frequent words, occurring over  
<sup>168</sup> 90% in a document, and the infrequent words, occurring under 0,05%, were kicked  
<sup>169</sup> out. The first dataset consists of editorial articles, editorial comments and forums.  
<sup>170</sup> The final number of documents and amount of words is listed in Table 3.1. The  
<sup>171</sup> second dataset consists of editorial articles and the summarized comments from the  
<sup>172</sup> editorials and forums. This is shown in Table 3.2. Both datasets were built for the  
<sup>173</sup> German and English language.

---

<sup>3</sup>contains the title, text and text of 100 comments

<sup>4</sup><https://radimrehurek.com/gensim/index.html>

<sup>5</sup><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.

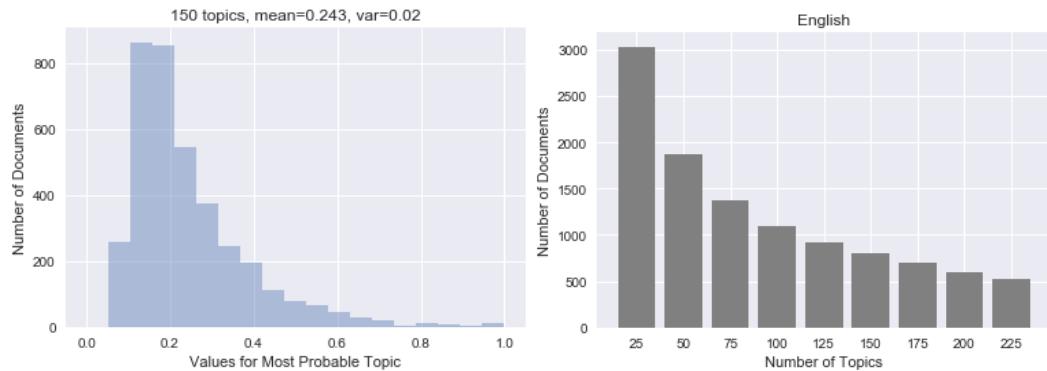
		Editorial articles	Comments
German	# documents	4730	2423
	# words	5239	22774
English	# documents	2345	3715
	# words	6254	17918

**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

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

- 189
- 190 1. For every Topic Model with different topic numbers a plot was generated, see  
 191 Figure 3.1. The x-axis shows the values for the most probable topic for every  
 192 single document while the y-axis shows the counted documents where the  
 193 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

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

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

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

200     After finding the appropriate topic number, the Topic Models generated with **NMF**  
201     and **LDA** for the same dataset were inspected manually. The domain experts labeled  
202     the topics and the Topic Model with the higher number of labels was chosen. The  
203     final selection of the Topic Models is shown in Table 3.3. And the Topic Models for  
204     the summarized comments is shown in Table 3.2.

205

	Editorial articles	Comments	Editorials		Forums
			articles	comments	
German	<b>190</b>	125	<b>190</b>	170	170
English	<b>130</b>	125	<b>130</b>	170	<b>110</b>

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

# 306 Experiments and Evaluation

## 208 4.1 Topic ranking

### 209 4.1.1 Related work

210 -Topic Significance Ranking of LDA Generative Models

### 211 4.1.2 Topic Coherence

212 Actually, good topics were determined from domain experts after topic labeling. The  
213 topics were labels from three labelers. If only one labeler or none of the labelers  
214 could assign a label the label was ranked with relevance score 0 otherwise with 1.

215 Actually, topic ranking with the help of labeling the topics

216 topics. However, not all the estimated topics are of equal importance or correspond  
217 to genuine themes of the domain. Some of the topics can be a collection of irrelevant  
218 or background words, or represent insignificant themes. identify and distinguish  
219 junk topics from legitimate ones, and to rank the topic significance. The

220 The setting of the number of latent variables K is extremely critical and directly  
221 effects the quality of the model and the interpretability of the estimated topics.  
222 Models with very few topics would result in broad topic definitions that could be  
223 a mixture of two or more distributions. topics is low in significance and often  
224 meaningless.

### 225 4.1.3 Summed Theta $\theta$

### 226 4.1.4 Inter-rater reliability

227 Fleiss' kappa (named after Joseph L. Fleiss) is a statistical measure for assessing the  
228 reliability of agreement between a fixed number of raters when assigning categorical  
229 ratings to a number of items or classifying items. cohens kapp only for max 2raters

<sup>230</sup> 0.01 – 0.20 Slight agreement <sup>231</sup> 0.21 – 0.40 Fair agreement <sup>232</sup> 0.41 – 0.60 Moderate agreement  
<sup>233</sup> 0.61 – 0.80 Substantial agreement <sup>234</sup> 0.81 – 1.00 Almost perfect agreement

<sup>232</sup> he value pj is the proportion of all assignments (raters \* number of topics) that were  
<sup>233</sup> made to the jth category.

<sup>234</sup> k:The value is between 0 and 1. The higher the value the better is the agreement  
<sup>235</sup> between the raters.

<sup>236</sup> N = anzahl an topics. n = anzahl an labelers (3) k = anzahl an labels

## <sup>237</sup> **Coherence**

<sup>238</sup> Hallo @Maria Potzner. Meinem Eindruck nach hat der Score wenig Aussagekraft.  
<sup>239</sup> Man kann kaum sagen, dass Topics, die weiter oben stehen "besser" sind als jene, die  
<sup>240</sup> weiter unten stehen. Einen Treshold festzulegen ist daher erst recht nicht möglich.

<sup>241</sup> SUMMED Theta:ab hier die topics rauswerfen", ein solcher Treshold wäre dann auch  
<sup>242</sup> bei jedem der 6 models unterschiedlich. Die topics am Ende kommen eben seltener  
<sup>243</sup> vor, es sind aber trotzdem sehr sinnvolle auch am Ende dabei, während am Anfang  
<sup>244</sup> auch einige "non-sense" topics stehen, zB Zeile 17 letztes Tabellenblatt.

<sup>245</sup> Also was für mich im Moment am besten erscheint wäre tatsächlich eine Selektion  
<sup>246</sup> und Labelling durch den Domain expert. Dann kann man z.B. noch ein Ranking  
<sup>247</sup> nach Topic-Häufigkeit machen.

## <sup>248</sup> 4.2 Automatic Topic Labeling

<sup>249</sup> Topic Models are used to discover latent topics in a corpus to help to understand  
<sup>250</sup> large collections of documents. These topics are multinomial distributions over all  
<sup>251</sup> words in a corpus. Normally, the top terms of the distribution are taken to represent  
<sup>252</sup> a topic, but these words are often not helpful to interpret the coherent meaning of  
<sup>253</sup> the topic. Especially, if the person is not familiar with the source collection. For  
<sup>254</sup> example, for the topic *price*, *\$*, *cost*, *foods*, *store*, *product*, *brand*, *low*, *supermarket*,  
<sup>255</sup> *good* a suitable label is *food prices*.

<sup>256</sup> With the help of Automatic Topic Labeling (**ATL**) we want to reduce the cognitive  
<sup>257</sup> overhead of interpreting these topics and, therefore, facilitate the interpretation of  
<sup>258</sup> the topics. Of course, the topics can be labeled manually by domain experts, but  
<sup>259</sup> this method is time consuming if there are hundreds of topics. Additionally, the  
<sup>260</sup> topic labels can be biased towards the users opinion and the results are hard to  
<sup>261</sup> replicate.

<sup>262</sup> We are working with domain specific data dealing with organic food. To generate  
<sup>263</sup> meaningful labels we can not make use of human turks because we need domain  
<sup>264</sup> experts who are proficient in this area. Therefore, we submitted the topics to our  
<sup>265</sup> domain experts to label them. But only 50 of the generated topics, ranked according  
<sup>266</sup> to the importance in a corpus, for each dataset were handed in, in order to not  
<sup>267</sup> burden them, since this process is very time-consuming. The datasets were labeled  
<sup>268</sup> by three labelers who tried to find a suitable label, which captures the meaning of  
<sup>269</sup> the topic and is easily understandable. After labeling, the three labels of a topic were  
<sup>270</sup> compared and a final label was set. If at least two labelers had the same label, this  
<sup>271</sup> was taken as the final one. If the given labels were not comparable, no label was set  
<sup>272</sup> at all.

<sup>273</sup> To relieve our domain experts in the following chapter two approaches for **ATL** are  
<sup>274</sup> described. In Section 4.2.2 an intrinsic method was used, which is only working  
<sup>275</sup> on texts and topics from our datasets to generate the labels according Mei et al.,  
<sup>276</sup> 2007. Section 4.2.3 describes an extrinsic approach by using a lexical database for  
<sup>277</sup> the English language called *Wordnet* to label the topics.

### <sup>278</sup> 4.2.1 Related work

<sup>279</sup> *Lau et al., 2011* generated a label set, called primary candidate labels, out of article  
<sup>280</sup> titles, which were found in Wikipedia or Google by searching the top N words from  
<sup>281</sup> 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.

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

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

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

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

346 *Mei et al., 2007* generated labels based on the texts collection and their related  
347 topics by chunking and building n-grams. They approximated the distribution for  
348 the labels and compared these to the distribution of the topic by calculating the  
349 Kullback Leibler (KL) divergence. To maximize the mutual information between  
350 the label and the topic distributions the calculated divergence has to be minimized.  
351 Three human assessors measured the results and found out that the final labels are  
352 effective and robust although applied on different genres of text collections.

## 353 4.2.2 Intrinsic Topic Labeling

354 The intrinsic topic labeling is based only on a text collection and therefrom extracted  
 355 topics. It does not use any external ontologies or embeddings. Because *Mei et al.*,  
 356 2007 were the only ones who generated topic labels by using an intrinsic approach,  
 357 we decided to apply their ATL on our data, using an implementation from Github<sup>1</sup>.  
 358 The implementation was adapted to our data and instead of using their preprocessing  
 359 ours was used.

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

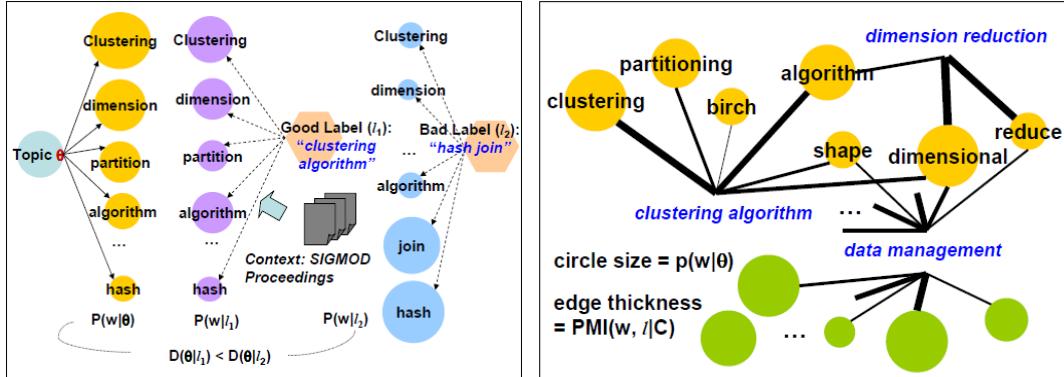
The candidate labels were ranked by their semantic similarity to the topic distribution  $\theta$ . To measure the semantic relevance between a topic and a label  $l$  a distribution of words  $w$  for the label  $p(w|l)$  was approximated by including a text collection  $C$  and a distribution  $p(w|l, C)$  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 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  $C$ :

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

365 We can see that the relevance scoring function consists of three parts. The first part  
 366 represents the expectation of PMI  $E_\theta(PMI(w, l|C))$  between  $l$  and the words in the  
 367 topic model given the context  $C$ , the second part is represented by the KL divergence

---

<sup>1</sup><https://github.com/xiaohan2012/chowmein>



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

368 between  $\theta$  and  $C$  and the third part can be viewed as a bias using context  $C$  to infer  
 369 the semantic relevance  $l$  and  $\theta$ . This bias can be neglected for our data because we  
 370 have used the same text collection for producing the topics and the labels. The same  
 371 applies to the second part, because the **KL** divergence has the same value for all  
 372 candidate labels. Therefore, we rank the topic labels with

$$Score(l, \theta) = E_\theta(PMI(w, l|C)) \quad (4.2)$$

373 The relevance scoring function is also described visually in Figure 4.1. The circles  
 374 represent the probability of terms. The larger the circle the higher is the probability.  
 375 On the left one can see that the label with lower **KL** divergence is the best one. To  
 376 approximate  $p(w|l)$  in this example the *SIGMOD Proceedings* were used as the text  
 377 collection  $C$ , not in our implementation. Analogously, we used our datasets. On  
 378 the right one can see a weighted graph, where each node is a term in the topic  
 379 distribution  $\theta$  and the edges between terms and the label are weighted by their **PMI**.  
 380 The weight of the node indicates the importance of a term to the topic, while the  
 381 weight of each edge indicates the semantical strength between label and term. The  
 382 relevance scoring function ranks a node higher if the label has a strong semantic  
 383 relation to the important topical words. Visually, this can be understood that the  
 384 label is ranked higher if it connects to large circle by a thick edge.

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

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

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

### 393 4.2.3 Extrinsic Labeling

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

400 We used the English online database *WordNet*<sup>2</sup>, that contains 118.000 different word  
401 forms and 90.000 word senses. WordNet organizes the several types of words like  
402 nouns, verbs, adjectives and adverbs into sets of synonyms, called *synsets*. A *synonym*  
403 is a word that has the same meaning as another word. E.g *shut* is a synonym for  
404 *close*. These two words form together with possibly other words such as *fold* a synset.  
405 Additionally, a synset contains a short definition, called *gloss*, and an exemplary  
406 sentence for each term in a synset, which describes the usage of this term. Every  
407 distinct word sense of a given word is represented as a separate synset. So the  
408 number of different meanings for a word corresponds to the number of synsets. All  
409 synsets are linked to each other according to semantic relations such as *synonymy*,  
410 *antonymy*, *hyponymy*, *hypernymy*, *meronymy* and *troponymy*. A definition of these  
411 semantic relationships can be found in *Miller, 1995*. In our implementation we used  
412 besides *synonymy* also *hypernymy*. If two words can be generalized by an other word,  
413 this word is called *hypernym*. E.g *animal* is a hypernym for *cat* and *dog*.

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

---

<sup>2</sup><http://wordnetweb.princeton.edu/perl/webwn>

Noun
<ul style="list-style-type: none"> <li>• <a href="#">S. (n) farming, agriculture, husbandry</a> (the practice of cultivating the land or raising stock)</li> <li>• <a href="#">S. (n) farming, land</a> (agriculture considered as an occupation or way of life) "farming is a strenuous life"; "there's no work on the land any more"</li> </ul>
Verb
<ul style="list-style-type: none"> <li>• <a href="#">S. (v) farm</a> (be a farmer; work as a farmer) "My son is farming in California"</li> <li>• <a href="#">S. (v) farm</a> (collect fees or profits)</li> <li>• <a href="#">S. (v) grow, raise, farm, produce</a> (cultivate by growing, often involving improvements by means of agricultural techniques) "The Bordeaux region produces great red wines"; "They produce good ham in Parma"; "We grow wheat here"; "We raise hogs here"</li> </ul>
Adjective
<ul style="list-style-type: none"> <li>• <a href="#">S. (adj) agrarian, agricultural, farming</a> (relating to farming or agriculture) "an agrarian (or agricultural) society"; "farming communities"</li> </ul>
Noun
<ul style="list-style-type: none"> <li>• <a href="#">S. (n) farming, agriculture, husbandry</a> (the practice of cultivating the land or raising stock) <ul style="list-style-type: none"> <li>◦ <a href="#">direct hyponym / full hyponym</a></li> <li>◦ <a href="#">part meronym</a></li> <li>◦ <a href="#">domain term category</a></li> <li>◦ <a href="#">direct hypernym / inherited hypernym / sister term</a></li> </ul> </li> <li>• <a href="#">S. (n) cultivation</a> ((agriculture) production of food by preparing the land to grow crops (especially on a large scale))</li> <li>• <a href="#">S. (n) production</a> ((economics) manufacturing or mining or growing something (usually in large quantities) for sale) "introduced more efficient methods of production"</li> <li>• <a href="#">S. (n) industry, manufacture</a> (the organized action of making of goods and services for sale) "American industry is making increased use of computers to control production"</li> <li>• <a href="#">S. (n) commercial enterprise, business enterprise, business</a> (the activity of providing goods and services involving financial and commercial and industrial aspects) "computers are now widely used in business"</li> <li>• <a href="#">S. (n) commerce, commercialism, mercantilism</a> (transactions (sales and purchases) having the objective of supplying commodities (goods and services))</li> <li>• <a href="#">S. (n) transaction, dealing, dealings</a> (the act of transacting within or between groups (as carrying on commercial activities)) "no transactions are possible without him"; "he has always been honest in his dealings with me"</li> </ul>

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

424 To extract the information from WordNet we used the *NLTK corpus reader*.<sup>3</sup> In  
 425 addition to WordNet also Polyglot<sup>4</sup> was used as kind of preprocessing for selecting  
 426 similar words of a topic by using word embeddings.

## 427 Preprocessing

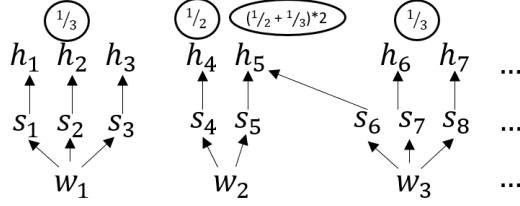
428 For all following approaches in the next section we implemented a preprocessing  
 429 step, that can be applied before running the different approaches for labeling a  
 430 topic. It should improve the quality of the labels. Our topics consists of 10 words,  
 431 usually these words can not be summarized to one label, which fits to all of the topic  
 432 words. Therefore, the distances between every combination of two topic words were  
 433 calculated with Polyglot embeddings. The top-5 words with the lowest distance  
 434 between each other were selected. On these top words the labeling methods were  
 435 applied.

## 436 Finding labels with a scoring function

437 Trying to find a good label for topics we used topic words  $w$  and generated synsets  $s$   
 438 for each topic word with the help of WordNet. Based on them we picked their direct  
 439 hypernyms  $h$ . To weight the hypernyms Custom scoring function ( $Csf$ ) was defined,  
 440 which includes the number of hypernyms  $h$  for the word  $w$  and the number of words,  
 441 that have a hypernym in common. When a hypernym for a word was found the  
 442 reciprocal of the total number of hypernyms for each word was assigned to to every  
 443 hypernym of the current word. If a selected hypernym is used by another word, too,

<sup>3</sup><http://www.nltk.org/howto/wordnet.html>

<sup>4</sup><https://polyglot.readthedocs.io/en/latest/Embeddings.html>



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

444 the scores are added and then multiplied by the number of common words. We  
 445 select the final label by the highest score.

446 Figure 4.3 illustrates the scores for each hypernym, which are represented as circles  
 447 above the hypernyms. The arrows connect the topic words  $w$  with their synsets  $s$ .  
 448 These are connected to hypernyms  $h$ . For  $w_1$  each hypernym  $h_1, h_2$  has the value  
 449  $\frac{1}{3}$ .  $h_4$  and  $h_5$  have the value  $\frac{1}{2}$ , but  $h_5$  is connected to  $s_5$  and  $s_6$ . Therefore, we add  
 450 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  
 451 highest score of  $\frac{5}{3}$  and is selected as the final label.

#### 452 **Find labels with similarity functions**

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

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

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

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

465  
 466 The remaining three similarity functions are measuring the Information Content (**IC**)  
 467 of synsets. **IC** combines the knowledge of the hierarchical structure from WordNet,  
 468 with statistics on actual usage in text as derived from a large corpus. Per default  
 469 WordNet uses the Brown Corpus. Although, this corpus is not related to our domain-

470 specific data, it includes a large number of English texts and is suitable as a reference  
471 corpus for this specific task.

- 472 • The *res-similarity* (Resnik-Similarity) weights edges between nodes by their  
473 frequency of the used textual corpus. Based on the **IC** of the Least Common  
474 Subsumer (lsc), the most specific ancestor node, a similarity score is calculated.
- 475 • The *jcn-similarity* (Jiang-Conrath Similarity) calculates the relationship between  
476 two words with  $(IC(w_1) + IC(w_2) - 2 * IC(lcs))$  and
- 477 • the *lin-similarity* calculates it with  $2 * IC(lcs)/(IC(w_1) + IC(w_2))$ .

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

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

#### 488 4.2.4 Evaluation

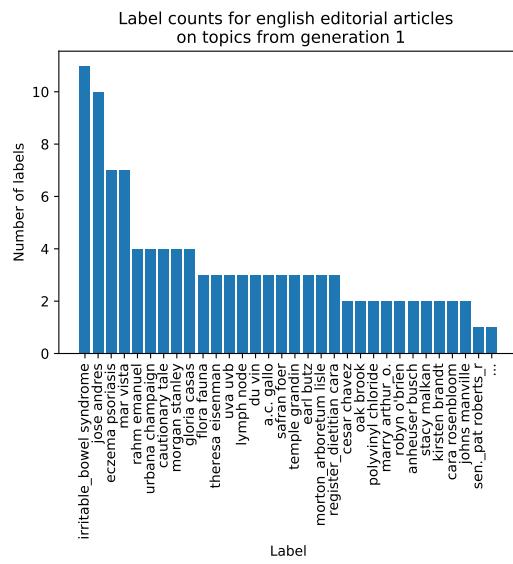
489 In the following section the results of intrinsic and extrinsic topic labeling will be  
490 evaluated regarding their quality and the number of different labels in a topic model.  
491 The labels generated automatically, are also compared to the manual labels, which  
492 were assigned by the domain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

##### 495 Intrinsic topic labeling

496 We applied the **ATL** in section 4.2.2 on our datasets, which include editorials, com-  
497 ments and forums. In general, the **ATL** according to Mei et al., 2007, outputs only  
498 different labels for topics, which were generated with **LDA**. For the topics generated  
499 with **NMF** the same label was given for every topic in a topic model. The reason

500 could be, that **NMF** does not return a probability distribution for every document.  
 501 Normalizing the values between 0 and 1 did not lead to an improvement. Therefore,  
 502 the labels for topics generated with **NMF** were neglected. Further evaluations are  
 503 based on English editorial articles.

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



[Fig. 4.4.:](#) Label counts for topics from Generation 1 according to Mei et al., [2007](#).

510

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

518 **Topics including POS-tagging:** By providing **POS**-tags, using Spacy<sup>5</sup>, we can limit  
 519 the labels to certain word types. In our experiments we used (NN-NN) or (JJ-NN)  
 520 **POS**-tags for English topic labels and (NN-NN) or (ADJD-NN) for German. To

<sup>5</sup>Possible POS-tags: <https://spacy.io/api/annotation>

	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 manual	rahm emanuel waste	hairy vetch 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 manual	irritable bowel syndrome dairy product	safran foer animal husbandry

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

521 apply **POS**-tagging, the preprocessing for the texts had to be changed, because in  
 522 Generation 1, a collocation finder was used. After performing this step the **POS**-tags  
 523 could not be applied retroactively. Therefore, we removed collocation finding and  
 524 added **POS**-tagging. All other preprocessing steps remained the same. Nevertheless,  
 525 the topics differ from the ones of Generation 1.

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

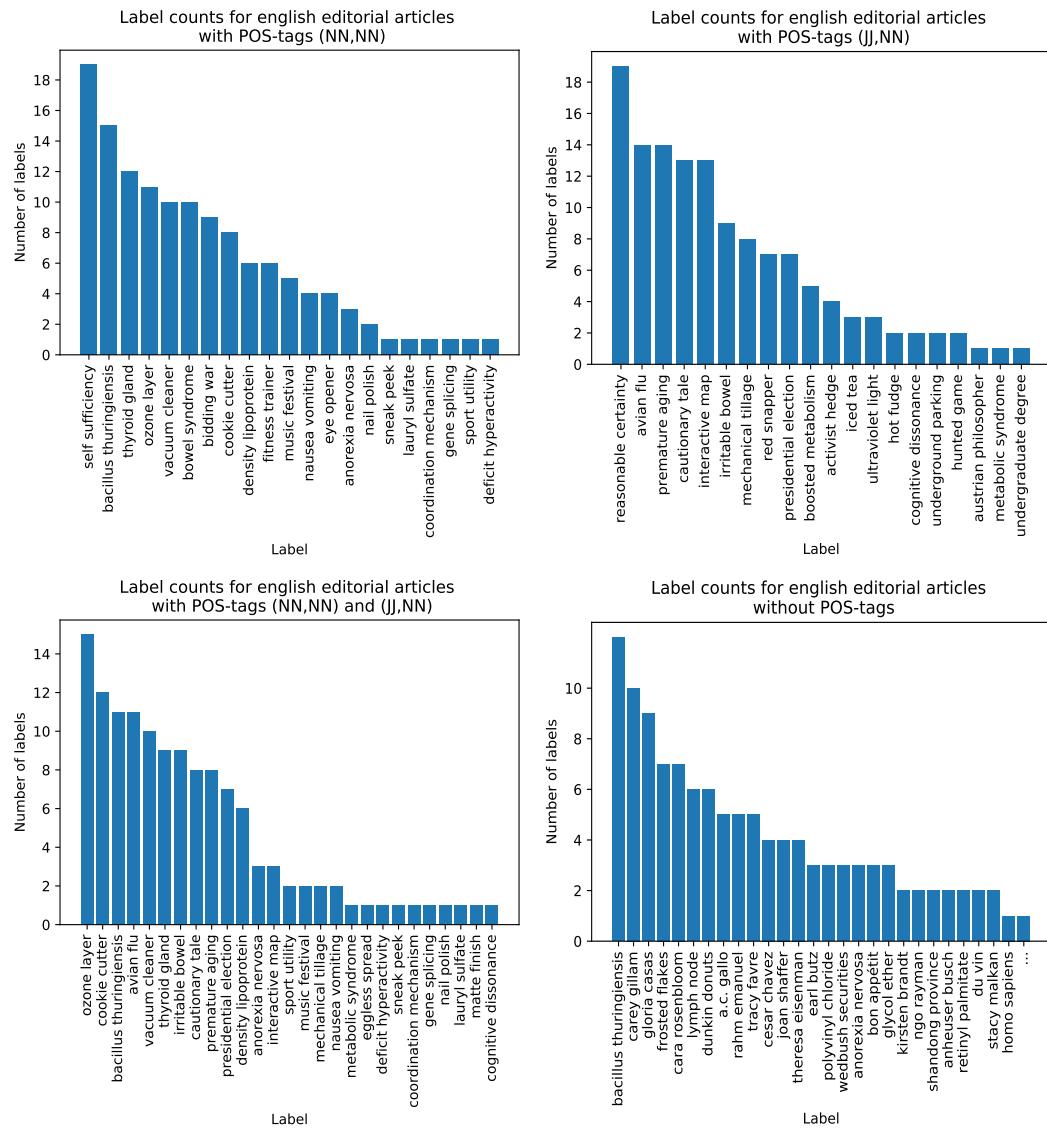
	Topic 6	Topic 10
with <b>POS</b> -tags	restaurant, fast, chain, meal, say, menu, ingredient, burger, chipotle, mcdonald	child, eat, kid, parent, family, healthy, school, who, health, can
(NN, NN) (JJ, NN) -	music festival hot fudge dunkin donuts	anorexia nervosa premature aging anorexia nervosa
	Topic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

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

530 In Figure 4.5 the label counts for English editorial articles using the texts, that were  
 531 **POS**-tagged are shown. On the x-axis all labels are listed, while the y-axis denotes  
 532 the number of same labels. In the plots where **POS**-tags were applied, no labels

533 include a name of persons and a smaller number of labels was outputted in contrast  
 534 to the plot without **POS**-tags.

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



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

542    **Extrinsic topic labeling**

543 Furthermore, we applied the ATL in section 4.2.3 on our Dataset, using the English  
544 online database WordNet and Polyglot embeddings. The described different similarity  
545 functions from WordNet, the Csf and the Polyglot embeddings were used to label our  
546 topics. A few examples are shown in Table 4.3 including the manual assigned labels  
547 to the topics, too. Some labels generated with the automatic approaches match  
548 the manual assigned labels. This is the case for the Topics 64, 84 and 107. For the  
549 other topics, the labels are heading to the same direction as the manual label: for  
550 Topic 97 *chemical* and manually *pesticide residues*, for Topic 99 *bee* and manually  
551 *beekeeping* and for Topic 109 *grocery store*, *mercantile establishment*, *marketplace*  
552 and manually *retailers* were assigned. Evaluating the automatically generated labels  
553 using different approaches, it was discovered that depending on the topics different  
554 labeling techniques output the best labels. It is not possible to tell, which approach  
555 is the best for all topics, let alone for several topic models according to the labels.  
556 Therefore, we tried to evaluate the labels generated with the extrinsic methods  
557 according to label counts. The words *entity*, *physical entity*, *object*, *whole*, *matter*  
558 and *abstraction* were chosen, because these are the most general words in the  
559 hierarchical tree of hypernyms in WordNet and do not have a high informative  
560 value. In Table 4.4 the number of non informative words are listed for the different  
561 similarity functions from WordNet. Based on the sum of the non informative words  
562 per similarity function and Polyglot embeddings (plg), we ranked the different  
563 methods in Table 4.5. The top 3 are: res-similarity with preprocessing, lin-similarity  
564 with preprocessing and Polyglot embeddings. The labels with Csf do not include  
565 any non informative words, because only the direct hypernyms and not the whole  
566 hierarchy of hypernyms were considered. Therefore, we plotted the amount of  
567 distinct labels in Figure 4.6. This shows, the labels generated with preprocessing on  
568 the left side and the labels without on the right. The number of same labels is at  
569 most 6 or 8, which shows that the labels are discriminative.

570 Having evaluated the intrinsic and extrinsic automatic topic labeling we can conclude,  
571 that the intrinsic approach generates meaningful and specific labels, that do not  
572 fit to the topics. The extrinsic approach generates partially good results, which  
573 are comparable with the labels from the domain experts. Nevertheless, finding  
574 meaningful and high qualitative labels is not yet automatable. The knowledge and  
575 experience a human person, which is required for topic labeling, can not be replaced  
576 by a machine.

	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 ich res jsn lin plg <b>Csf</b> manual	entity entity produce produce produce vegetable cultivate homegrown food	produce produce produce produce produce vegetable cultivate food	abstraction abstraction <b>dairy product</b> produce beverage <b>dairy product</b> nakedness <b>dairy product</b>	beverage produce beverage beverage beverage abstraction farm
	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 ich res jsn lin plg <b>Csf</b> manual	entity entity matter food matter cattle cattle animal husbandry	meat abstraction meat meat meat physical entity be husbandry	beverage physical entity substance substance <b>beverage</b> <b>beverage</b> food <b>beverage</b> <b>beverage</b>	beverage substance substance <b>beverage</b> <b>beverage</b> food <b>beverage</b>
	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 ich res jsn lin plg <b>Csf</b> manual	matter matter matter matter produce fruit chemical pesticide residues	matter matter matter matter matter entity chemical residues	organism organism organism organism bee bee farmer beekeeping	person person organism whole artifact artifact scientist
	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 ich res jsn lin plg <b>Csf</b> manual	material abstraction material abstraction material <b>waste</b> convent <b>waste</b>	material physical entity material material material abstraction lowland	grocery store physic entity social group grocery store social group artifact marketplace retailer	mercantile establishment mercantile establishment mercantile establishment mercantile establishment mercantile establishment abstraction marketplace

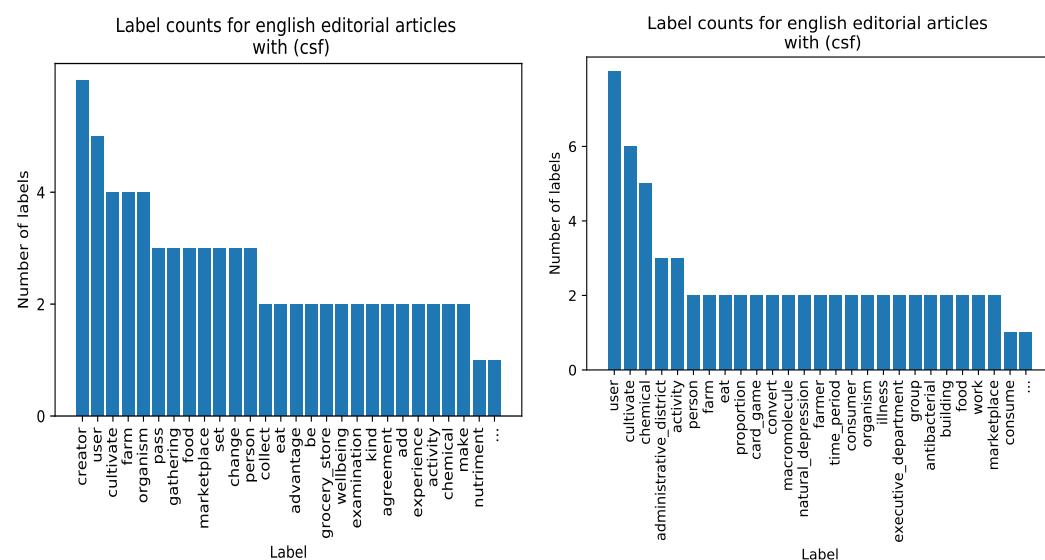
**Tab. 4.3.:** Topics labeled from Generation 1 manually and with extrinsic methods. Labels including preprocessing are in the third and fifth column. **Bold** words are the same as the manual assigned label.

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

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

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

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



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

## 577 4.3 Internal consistency

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

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

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

604 For our evaluation we generated for every dataset topic models with 25, 50, 75  
605 topics and topic models with 50 topics over and under the topics number from  
606 *Generation 1*. The different key figures were applied on these newly generated topic  
607 models and the topic models from *Generation 1*. In the following the dataset for  
608 German editorial articles with 25, 50, 75, 140, 190 and 240 topics per topic model  
609 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

### 610 4.3.1 Theta $\theta$

611 The document topic matrix  $\theta$  describes to which extend a topic is represented in a  
 612 certain document. We used the matrix to calculate the number of documents, which  
 613 are assigned to a topic and the number of topics, which are assigned to a document.  
 614 In both cases a threshold fo 10% was used. The example document topic matrix  
 615 in Table 4.6 shows two documents and 5 topics. For topic 1 the counter number of  
 616 documents is only 1 (Document 1). For Document 1 the relevant number of topics is  
 617 4 (Topic 1,2,4,5).

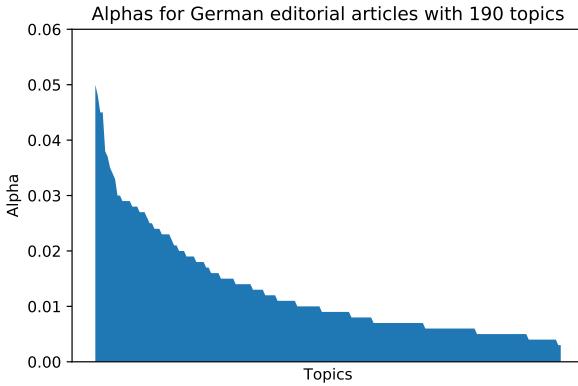
### 618 4.3.2 Alpha

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

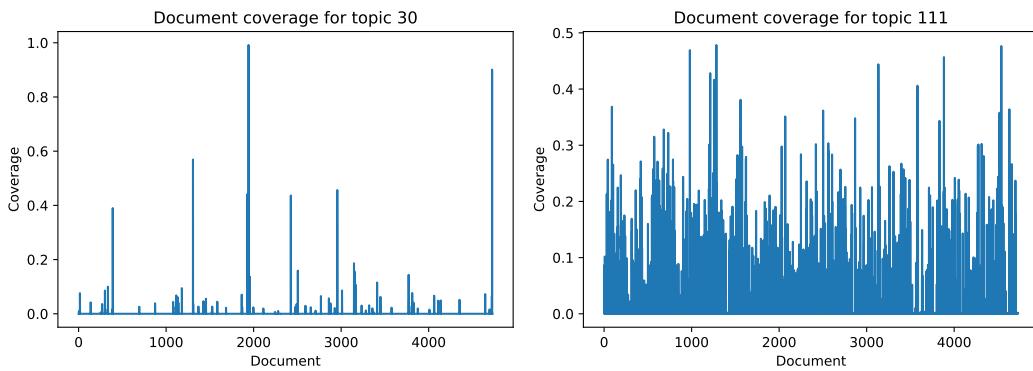
628 In Figure 4.7a the alpha values for each topic for German editorial articles with 190  
 629 topics are shown. The document distribution for Topic 30, which has the lowest  
 630 alpha value, is considered in Figure 4.7b. The x-axis represents the document ids,  
 631 standing for the documents, which build the corpus. The y-axis represents the  
 632 percentage of a document that is covered by the topic. In contrast to Topic 30, Topic  
 633 111 with the highest alpha value is shown in Figure 4.7c. One can see, that Topic  
 634 30 covers only a few documents, while Topic 111 is more evenly spread over all  
 635 documents.

### 636 4.3.3 Entropy

637 Entropy was used to identify specific and general topics. It can be applied on the  
 638 topic term matrix  $\phi$  and the document topic matrix  $\theta$ . When applied on the topic



(a) Plotted alphas for German editorial articles



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

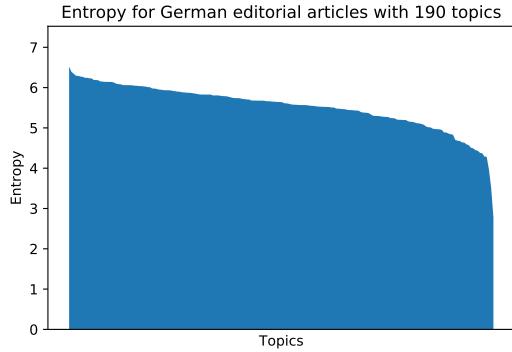
(c) Document coverage for the topic with the highest alpha value

**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

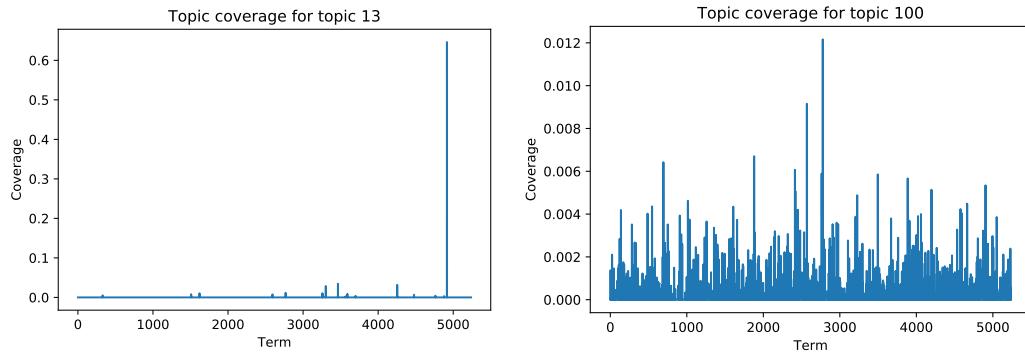
term matrix, a high entropy value indicates that the topic is rather general. This means, that all terms have a similar probability to appear in the topic. A low entropy indicates, that the topic is specific i.r. only a few words have a high probability to appear in the topic. This difference is illustrated 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):

$$E = - \sum (p * \log(p)) \quad (4.4)$$

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



(a) Plotted entropy for German editorial articles



(b) Topic coverage for the topic with the lowest entropy      (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

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.

#### 4.3.4 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:

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

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

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

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

### 679 4.3.5 Jensen Shannon divergence

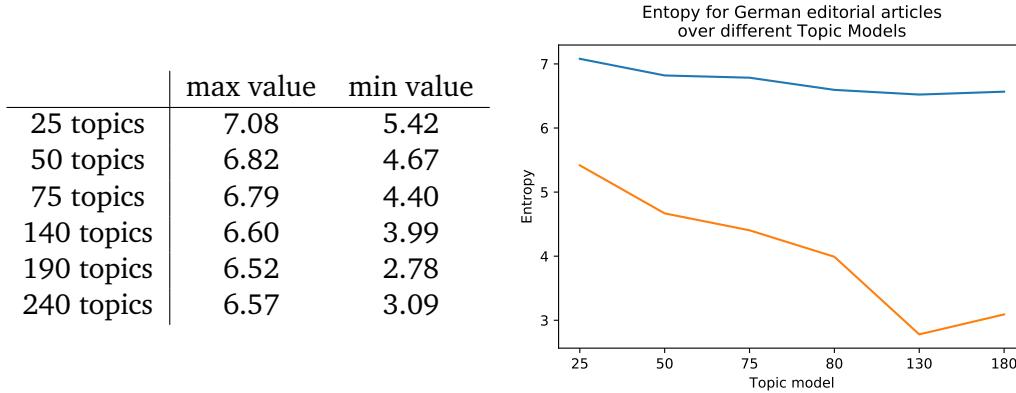
680 The topics returned by LDA are probability distributions over all terms in the corpus.  
 681 Therefore, to compare the similarity of two topics  $p$  and  $q$  we can use existing  
 682 metrics to measure the similarity between probability distribution. Lin, 1991 et al  
 683 lists possible similarity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

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

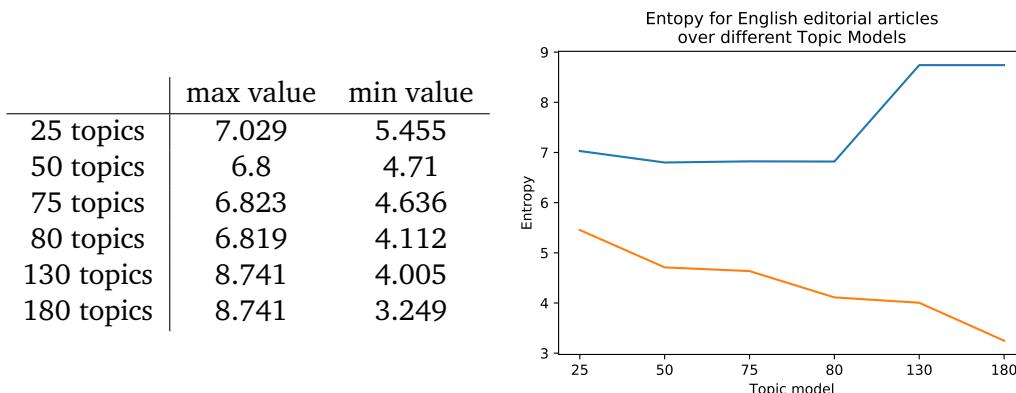
686 where  $j$  is the number of a certain term and  $T$  describes the total number of terms.  
 687  $p_j$  represents the probability of term  $j$  appearing in topic  $p$ .  $q_j$  is the probability  
 688 of term  $j$  appearing in topic  $q$ . The KL divergence is an asymmetric measurement.  
 689 For our use-case, comparing the topics intra and inter a topic model, a symmetric  
 690 measure is needed, which guarantees the same results for the comparison of  $t_1$  with  
 691  $t_2$  and  $t_2$  with  $t_1$ . Therefore, based on the KL divergence, the Jensen Shannon (JS)  
 692 divergence is used:

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

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



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



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

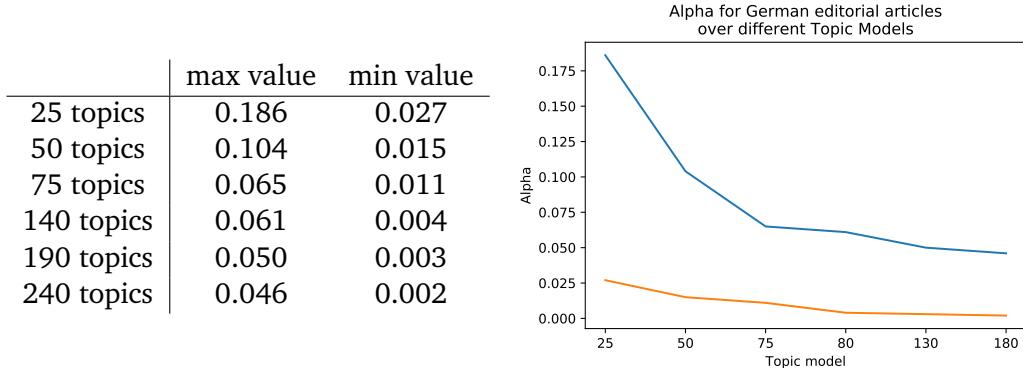
698 two probability distributions are identical and the value 0 when they are completely  
699 dissimilar.

### 700 4.3.6 Evaluation

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

#### 707 Entropy

708 The Figures 4.9 and 4.10 show the change of entropy for German respectively English  
709 editorial articles. The table on the left denotes the minimal and maximal entropy



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

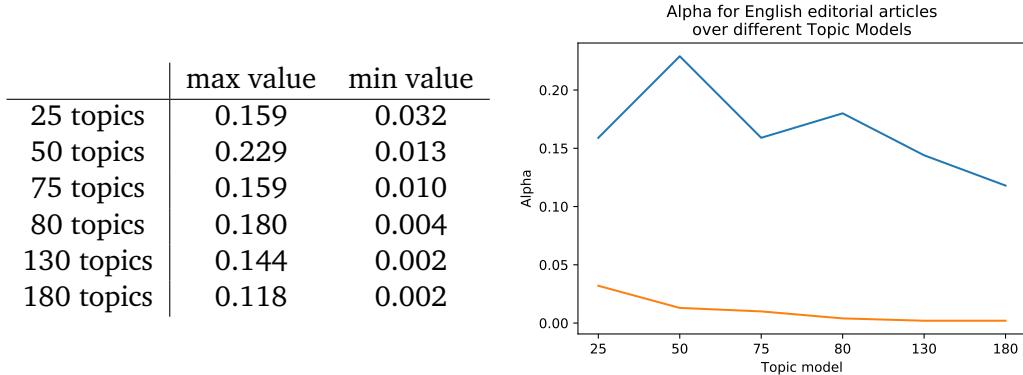
given the number of topics. On the right the maximal entropy values (blue line) and the minimal entropy values (orange line) are plotted. This structure will repeat in the different key figures.

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.

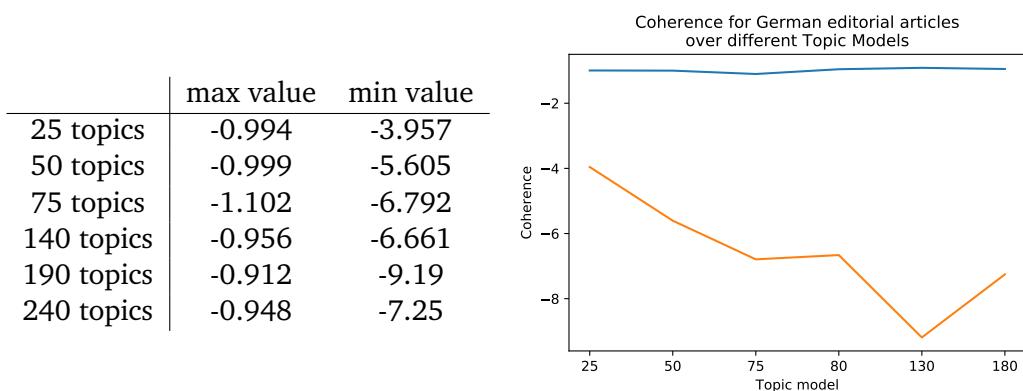
For *English editorial articles* (Figure 4.10) the maximal entropy value is decreasing up to the topic model with 80 topics. Then the values is rising and for the topic models with 130 and 180 topics the entropy stays the same. For the minimal entropy the values are getting continuously smaller. So the 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.

## 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



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



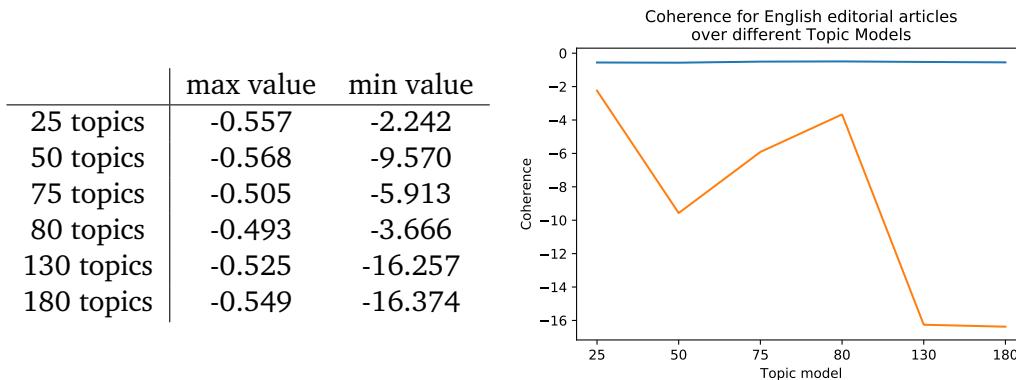
**Fig. 4.13.:** Maximal and minimal coherence values per topic model for German editorial articles.

735 alpha value consists out of either general topics and the document with the lowest  
 736 alpha value out of specific ones. (!!!stimmt nur für die beiden, die entropy über alle  
 737 alpha values sind sprunghaft...!!!)

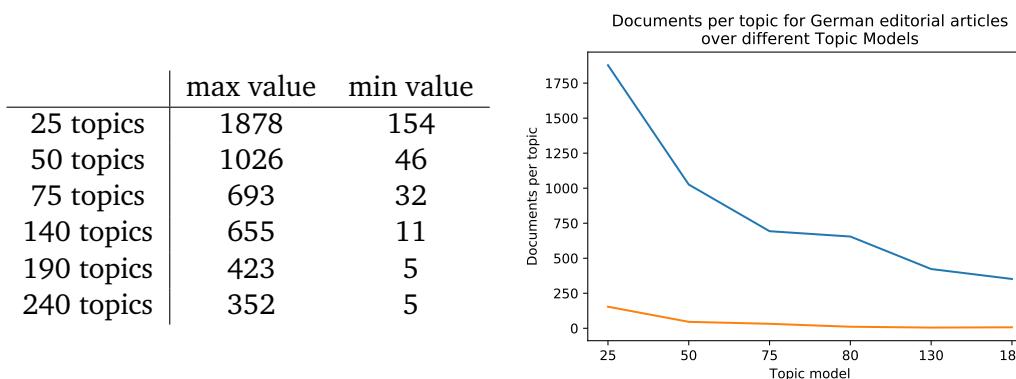
738 The minimal alpha values for *English editorial articles* (Figure 4.12) are continuously  
 739 declining up to the topic model with 130 topics. Then the alpha value is staying the  
 740 same. The maximal alpha values are volatile, so there is no prediction possible, if  
 741 the values will raise or fall again.

## 742 Coherence

743 The maximal coherence score is for *German editorial articles* (Figure 4.13) in the  
 744 range of the maximal values -0.91 and -1.1 and for *English editorial articles* (Figure  
 745 4.14) in the range of -0.49 and -0.56. The maximal values do not change a lot for  
 746 both datasets, but no pattern how the values are changing can be seen. The same  
 747 can be said for the minimal coherence values, the only difference is, that the range  
 748 in which the coherence is moving is much bigger than the range from the maximal



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



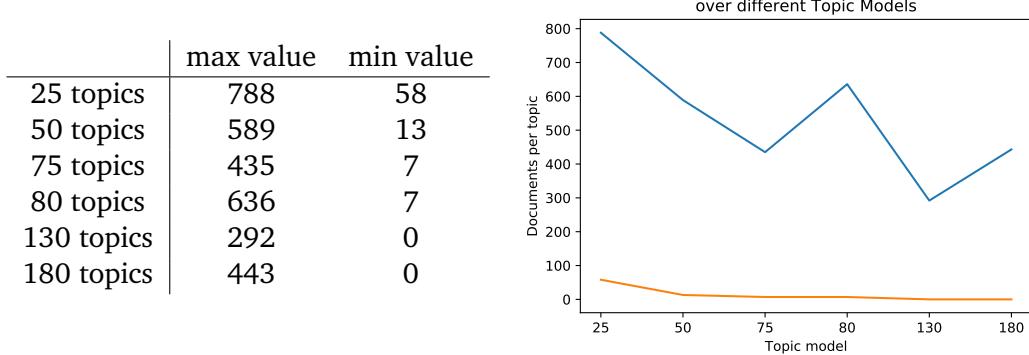
**Fig. 4.15.:** Maximal and minimal number of documents containing the same topic for German editorial articles.

749 values. For *German editorials* it is between -3.96 and -9.2 and for *English editorials*  
 750 between -2.2 and -16.4.

## 751 Theta

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

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

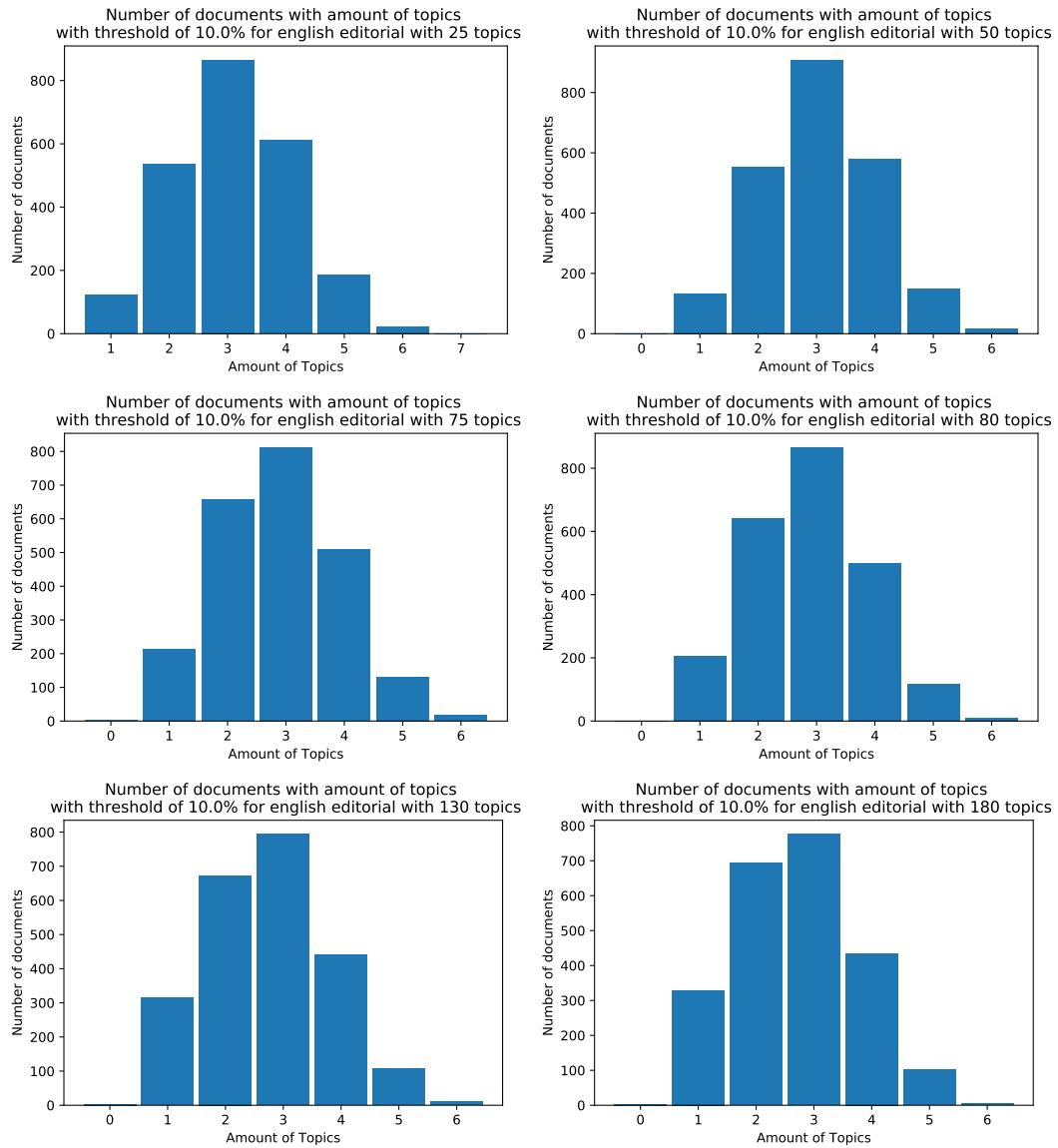


**Fig. 4.16.:** Maximal and minimal number of documents containing the same topic for English editorial articles.

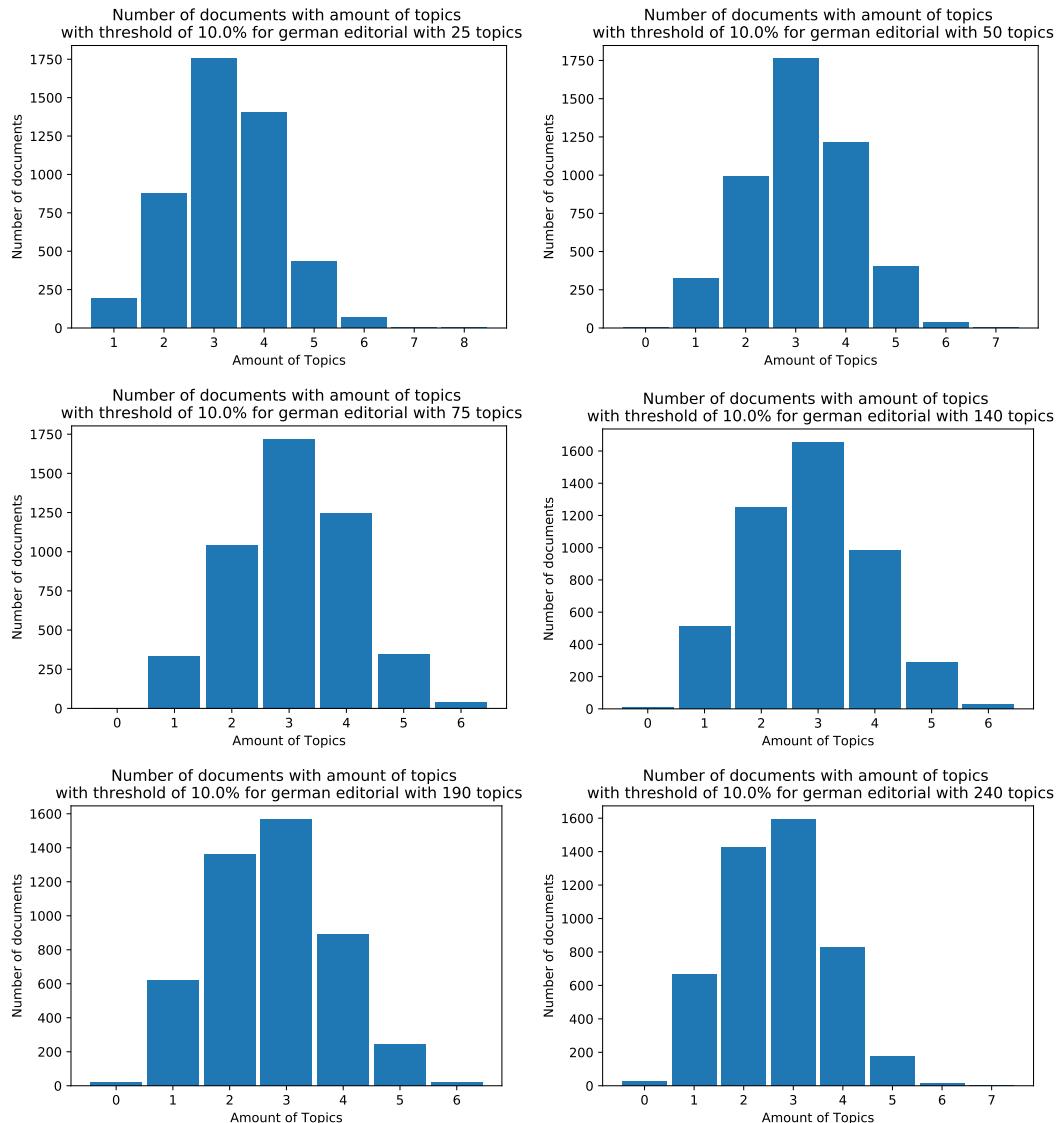
760 For *English editorial articles* (Figure 4.16) the maximal number of documents is  
 761 volatile, whereas the minimal numbers of documents are decreasing and there even  
 762 seem to be topics, that are not assigned to any document, because they do not cover  
 763 the meaning of the document over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

764 Second, we analyzed how many topics occur in a document over 10%. The plots in  
 765 Figure 4.18 and 4.17 represent the number of documents with the number of topics  
 766 over the threshold. On the x-axis the amount of topics, which are occurring over  
 767 10% in a document, while the y-axis represents the number of documents, which  
 768 include a certain amount of topics.

769 In Figure 4.17 is shown how the number of topics for *English editorial articles* change.  
 770 When increasing the number from 25 to 50 topics, all documents are modeled  
 771 by a most 6 topics. This means, that with a higher topic number the topics are  
 772 more tailored to documents and thus the documents can be represented with less  
 773 topics. This is also supported by the observation, that the number of documents, that  
 774 contain only one or two topics over the threshold are slightly increasing with the  
 775 number of topics. At the same time there are some documents, that do not express  
 776 any topic over the threshold, which could indicate, that these documents are rather  
 777 general and cover many topics with a low probability, instead of covering a few  
 778 specific topics. For the *German editorial articles* in Figure 4.18 it can be seen, that the  
 779 maximal number of topics, covering a document is also at 3. From the topic model  
 780 with 25 topics to 50 topics, there are no documents, which include 8 topics, but  
 781 there are documents, that do not contain any topic. When increasing the number  
 782 of topics to the next level, there are no documents, which contain 7 topics. In the  
 783 next plots, the number of documents containing 2, only 1 or no topic is raising, but  
 784 the maximal topic number remains 3. The last plot is looking nearly the same apart  
 785 from adding a few documents with 7 topics. For this dataset it could be said, that  
 786 the topics are getting up to 190 topics more specific. With more topics the number



**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

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

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

787 of generic topics increases. So the topic model with 190 seems the best for obtaining  
 788 good topics.

### 789 Jensen Shannon divergence

790 The **JS** divergence was used to calculate the similarity of topics based on the topic  
 791 term matrix in a topic model and across topic models. The similarities were then  
 792 visualized as heat maps. The dark red colored boxes express high similarity, while the  
 793 brighter boxes express a smaller one. The topic models form *German editorial articles*  
 794 with 25, 50 and 70 topics were evaluated manually. The highest similarities (dark  
 795 red) with the values of 0.63, 0.65, and 0.65 for the topics intra a topic model were  
 796 taken, and the topics were compared with each other in Table 4.7. The same words  
 797 form the top 10 terms of a topic were marked in bold. In Table 4.8 the similarities  
 798 of the topics were calculated across different topic models with 25/50, 50/57 and  
 799 50/75 topics. The values for the similarity were 0.9, 0.87 and 0.94. One can see,  
 800 that the number of common words and the calculated similarity, is much higher in  
 801 the inter topic model evaluation. This means, that the topics inter a topic model are  
 802 not as similar to each other as the topics developing across topic models. At least  
 803 this is not recognizable by the top 10 words of a topic. The heat maps for the inter  
 804 topic model and intra topic model evaluation can be seen in Figure ?? and 4.21.

### 805 Correlation

806 -tried different correlations, alpha and entropy correlation depends on the topic  
 807 models (num of topics)

808 -alpha und amount doc per topic correliert zu 100 immer

topics	compared topics with the highest similarity:	
25/50	T21: <b>prozent, euro, ökologisch, betrieb,</b> hektar, million, steigen, fläche, deutschland, zahlen	T2: <b>prozent, ökologisch, betrieb, hektar,</b> fläche, landwirtschaft, zahlen, anteil, bewirtschaften, euro
50/75	T4: <b>pestizid, finden, probe, rückstand,</b> greenpeace, konventionell, untersuchen, belasten, prozent, Einsatz	T54: <b>pestizid, rückstand, probe, grenzwert,</b> finden, greenpeace, stoff, belasten, untersuchen, einsetzen
50/75	T9: <b>eiern, fipronil, belasten, niederlande,</b> deutschland, nehmen, verkaufen, betreffen, betrieb, angeben	T57: <b>eiern, fipronil, belasten, niederlande,</b> 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	25/50	0.35	0.91
25	0.38	0.63	25/75	0.35	0.88
50	0.35	0.65	50/75	0.35	0.34
75	0.33	0.65	140/190	0.33	0.88
140	0.34	0.62	140/240	0.32	0.92
190	0.33	0.61	190/240	0.33	0.91
240	0.33	0.60			

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

(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	topics	min value	max value
25	0.42	0.66	25/50	0.39	0.90
50	0.36	0.68	25/75	0.37	0.91
75	0.35	0.69	50/75	0.35	0.91
80	0.35	0.67	80/130	0.34	0.95
130	0.34	0.65	80/180	0.33	0.97
180	0.36	0.67	130/180	0.33	1.0

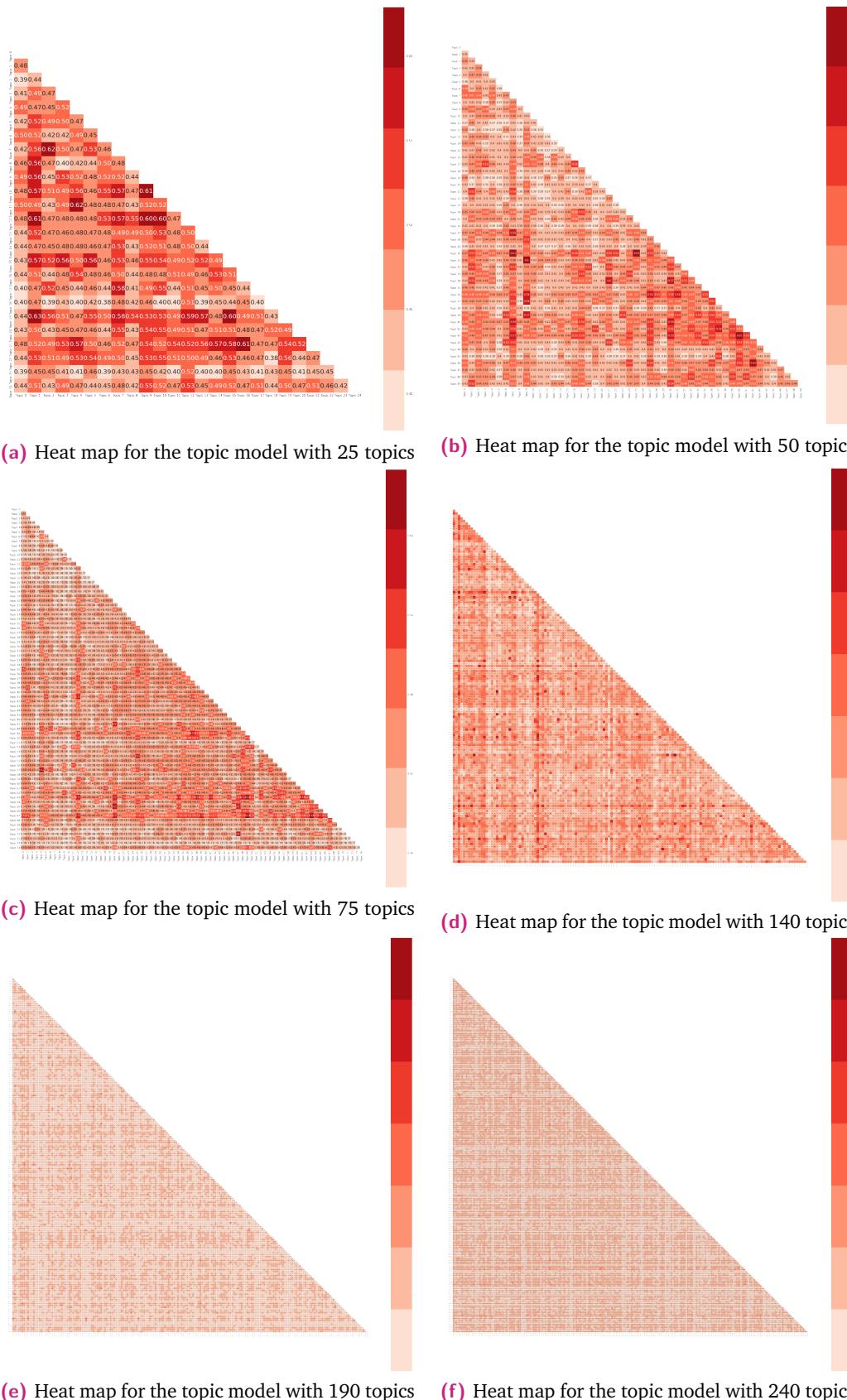
(a) Minimal and maximal similarities for English editorial articles inter different topic models

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

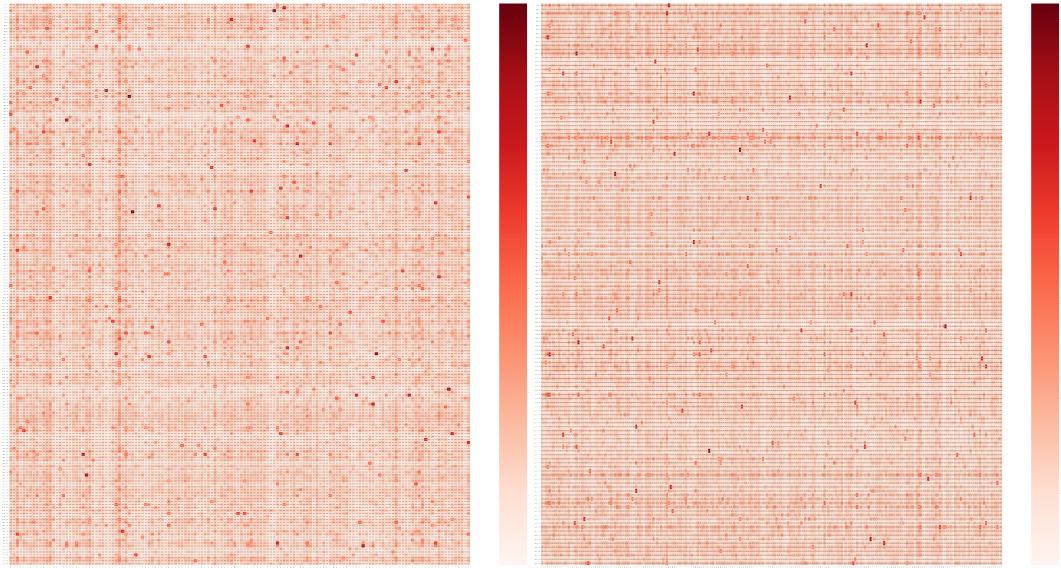
**Fig. 4.20.:** Minimal and maximal similarities inter and intra topic models for English Editorial articles

topics	min value	max value
25/50	0.35	0.91
25/75	0.35	0.88
50/75	0.35	0.34
140/190	0.33	0.88
140/240	0.32	0.92
190/240	0.33	0.91

**Tab. 4.9.:** Minimal and maximal similarities for German editorial articles inter different topic models



**Fig. 4.21.: Heat maps with similarities inter a topic models for German editorial articles**



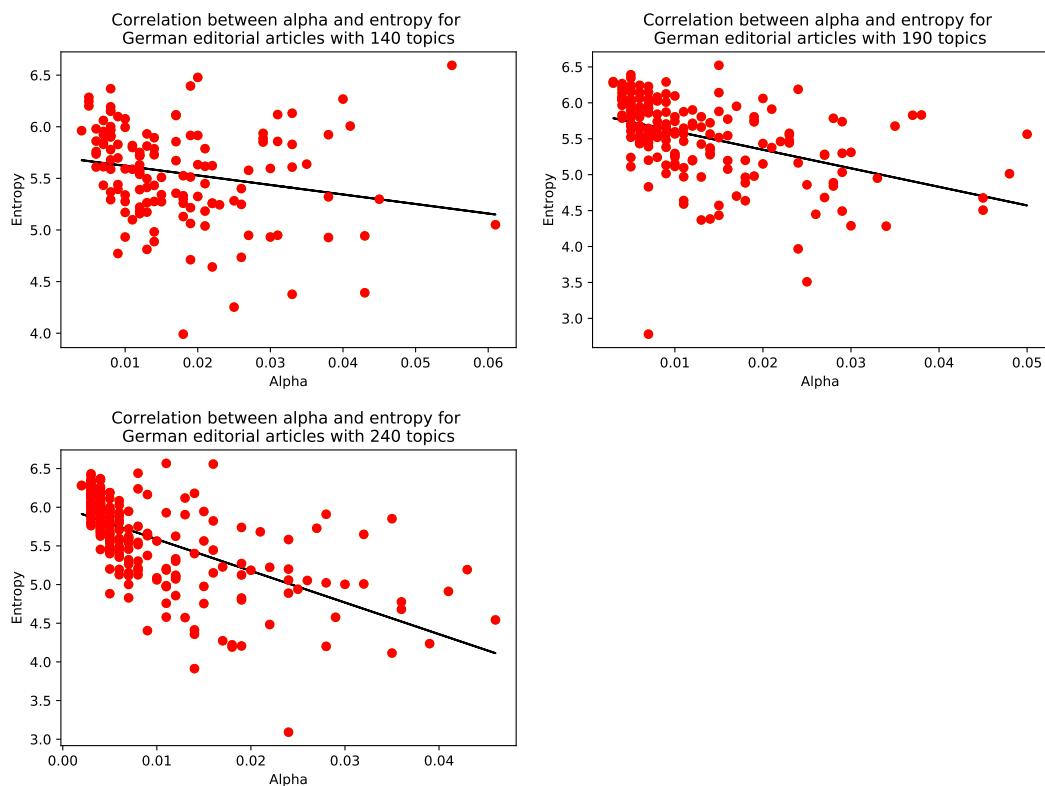
(a) Heat map for the topic models with 140 and 190 topics (b) Heat map for the topic models with 140 and 240 topics



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

**Fig. 4.22.:** Heat maps with similarities intra two topic models for German editorial articles

809 soll ich die anderen kennzahlen, auch noch auf korrelation untersuchen?



**Fig. 4.23.: ea**

[Correlation between alpha and entropy for German editorial articles]

810 pearson correlation: 80: 0.18203719 130: -0.18759358 180: -0.24035685

811 resume: besten kriterien lassen sich nicht so leich bestimmen. da es vom datensatz  
812 abhängen kann, aus wie vielen dokumenten besteht der corpus, wie lang sind diese  
813 documente. und wie spezifisch sind die themen, die in einem document behandelt  
814 werden.

# 815 Future Work and Conclusion

## 817 5.1 Future work

818 - Mai bessere Kl oder andere funktion generieren, sodass die labels besser zu den  
819 topics passen

820 - haben topic labeling mit ontology nur auf englischsprachigen topics angewandt,  
821 man könnte versuchen, dies auch mit Thesaurus db auf den deutschen topics anzu-  
822 passen.

823 - ein verfahren für topic labeling, dass man unabhängig von der sprache bzw vom  
824 algorithmus lda bzw nmf und weitere, verwenden kann.

825 Intern consistency: compare topic models based on changing other parameters such  
826 as alpha, beta, for lda, nmf only topic number as übergabeparameter

## 827 5.2 Conclusion

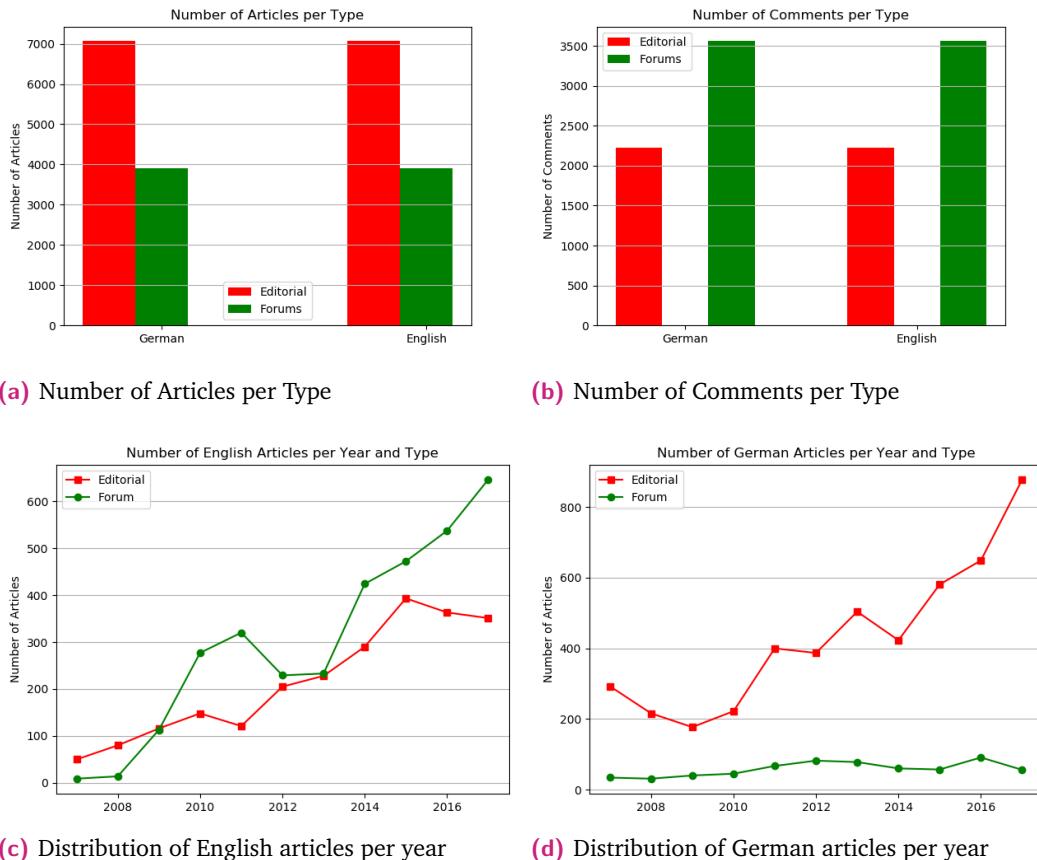
828 In this thesis two main topics were covered: How can we label topics automatically  
829 and how can we measure the internal consistency of topic modeling.

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

837 To answer the second question, several key figures were introduced to measure the  
838 internal consistency of topic models with a different number of topic numbers. All  
839 key figures were applied on our dataset.

<sup>840</sup> ... No of the key figures is able to answer the question. Further we also did  
<sup>841</sup> correlations....

## 843 Descriptive Statistics of the Dataset



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

### 844 A.1 Detailed Statistics of all Sources

### 845 A.2 JSON Storage Schema

<sup>1</sup>The average number of tokens after lemmatizing and stop word removal.

Source	Total articles	Relevant articles	% rel. articles	Avg. article length <sup>1</sup>	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 <sup>1</sup>
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 <sup>1</sup>	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

Source	Total comments	Relevant comments	% rel. cmnt.	Root cmnt.	% root cmnt.	Avg. # cmnt.	Avg. cmnt. length <sup>1</sup>
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.3:** Article statistics for English forum data

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

Source	Total articles	Relevant articles	% rel. articles	Avg. article length <sup>1</sup>	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 <sup>1</sup>
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	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
bioekoforum	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 <sup>1</sup>
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
bioekoforum	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

```

1   {
2     "article_title": "article title",
3     "article_author": [
4       {
5         "article_author_id": "123456789",
6         "article_author_name": "author name"
7       }
8     ],
9     "article_time": "2015-10-17 20:02:54",
10    "article_text": "article text",
11    "article_source": "news source",
12    "comments": [
13      {
14        "comment_id": "123456789",
15        "comment_author": {
16          "comment_author_id": "45678",
17          "comment_author_name": "author name",
18        },
19        "comment_time": "2015-10-20 04:17:17",
20        "comment_text": "comment text",
21        "comment_rating": -15.0,
22        "comment_title": "example title"
23      },
24      {
25        "comment_id": "987654321",
26        "comment_author": {
27          "comment_author_id": "12345",
28          "comment_author_name": "author name"
29        },
30        "comment_time": "2015-10-19 19:16:33",
31        "comment_text": "comment text",
32        "comment_replyTo": "123456789",
33        "comment_rating": 6.0
34      }
35    ],
36    "search_query": "organic farming",
37    "article_url": "https://example.url",
38    "resource_type": "editorial | blog | forum",
39    "article_rating": 5.0
40  }

```

**Listing 1:** JSON Storage Schema

# <sup>846</sup> Bibliography

- <sup>847</sup> AGOF (2018). *Nettoreichweite der Top 15 Nachrichtenseiten (ab 14 Jahre) im November 2014 in Unique Usern (in Millionen)* (cit. on p. 7).
- <sup>849</sup> Allahyari, Mehdi and Krys Kochut (2015). „Automatic Topic Labeling using Ontology-based Topic Models“. In: (cit. on p. 14).
- <sup>851</sup> Ankit Sethi, Bharat Upadrasta (2012). „Introduction to Probabilistic Topic Modeling“. In: 151, pp. 10–17 (cit. on p. 30).
- <sup>853</sup> Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin (2016). „Automatic Labelling of Topics with Neural Embeddings“. In: 1, pp. 953–963. arXiv: [1612.05340](https://arxiv.org/abs/1612.05340) (cit. on pp. 15, 18).
- <sup>855</sup> Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). „Latent Dirichlet Allocation“. In: *Journal of Machine Learning Research* 3.3/1/2003, pp. 993–1022. arXiv: [1111.6189v1](https://arxiv.org/abs/1111.6189v1) (cit. on p. 29).
- <sup>858</sup> 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. 15, 18).
- <sup>861</sup> IVW (2018). *Verkaufte Auflage der überregionalen Tageszeitungen in Deutschland im 3. Quartal 2018* (cit. on p. 7).
- <sup>863</sup> 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](https://arxiv.org/abs/1011.1669v3) (cit. on p. 5).
- <sup>866</sup> 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. 15).
- <sup>870</sup> 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. 13, 15, 18).
- <sup>873</sup> Lin, Jianhua (1991). „Divergence Measures Based on the Shannon Entropy“. In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151 (cit. on p. 32).
- <sup>875</sup> 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. 15, 18).

- 878 Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to*  
879 *Information Retrieval*. arXiv: [052186571](https://arxiv.org/abs/052186571) (cit. on p. 4).
- 880 Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). „Automatic labeling of multinomial  
881 topic models“. In: *Proceedings of the 13th ACM SIGKDD international conference on*  
882 *Knowledge discovery and data mining - KDD '07* January 2007, p. 490 (cit. on pp. 13–17,  
883 21, 22).
- 884 Miller, George A. (1995). „WordNet: a lexical database for English“. In: *Communications of*  
885 *the ACM* 38.11, pp. 39–41 (cit. on p. 18).
- 886 Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum  
887 (2011). „Optimizing Semantic Coherence in Topic Models“. In: *Proc. of the Conference*  
888 *on Empirical Methods in Natural Language Processing, EMNLP '11* 2, pp. 262–272 (cit. on  
889 p. 31).
- 890 Newman, David, Timothy Baldwin, and Edmund Talley (2010). „Evaluating Topic Models  
891 for Digital Libraries“. In: (cit. on p. 31).
- 892 Salton, G, A Wong, and C S Yang (1975). „1975.A vector space model for automatic indexing.pdf“.  
893 In: 18.11 (cit. on p. 5).
- 894 Stevens, Keith;Kegelmeyer,Philip;Andrzejewski, David;Buttler, David (2012). „Exploring  
895 Topic Coherence over many models and many topics“. In: (cit. on pp. 31, 32).
- 896 Steyvers, M and T Griffiths (2007). „Probabilistic topic models“. In: *Handbook of Latent*  
897 *Semantic Analysis: A Road to Meaning*, pp. 424–440. arXiv: [1111.6189v1](https://arxiv.org/abs/1111.6189v1) (cit. on pp. 29,  
898 32).
- 899 Widmer, Christian (2018). „Topic Modeling for Opinion Mining“. In: (cit. on p. 2).
- 900 Zhao, Wayne Xin, Jing Jiang, Jing He, et al. (2011). „Topical keyphrase extraction from  
901 Twitter“. In: *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Com-*  
902 *putational Linguistics: Human Language Technologies - Volume 1*, pp. 379–388 (cit. on  
903 p. 14).

# <sup>904</sup> List of Figures

905 <b>3.1</b>	Count of the value of the most probable topic, summed over all topics.	10
906 <b>3.2</b>	Number of documents the topics are expressed above the threshold . . . . .	10
907 <b>4.1</b>	Relevance scoring function for ATL . . . . .	17
908 <b>4.2</b>	WordNet results for the word <i>farming</i> . . . . .	19
909 <b>4.3</b>	ATL: Scoring function for hypernyms . . . . .	20
910 <b>4.4</b>	Label counts for topics from Generation 1 with intrinsic labeling . . . . .	22
911 <b>4.5</b>	Label counts for topics including POS-tags with intrinsic method. . . . .	24
912 <b>4.6</b>	Label counts for topics from Generation 1 with Csf. . . . .	27
913 <b>4.9</b>	Maximal and minimal entropy per topic model for German editorial articles. . . . .	33
915 <b>4.10</b>	Maximal and minimal entropy values per topic model for English editorial articles. . . . .	33
917 <b>4.11</b>	Maximal and minimal alpha values per topic model for German editorial articles. . . . .	34
919 <b>4.12</b>	Maximal and minimal alpha values per topic model for English editorial articles. . . . .	35
921 <b>4.17</b>	Amount of topics in documents over a threshold of 10% for English editorial articles . . . . .	38
923 <b>4.20</b>	Minimal and maximal similarities inter and intra topic models for English Editorial articles . . . . .	41
925 <b>4.23</b>	ea . . . . .	44
926 <b>A.1</b>	Descriptive Statistics for all datasets . . . . .	47

# <sup>927</sup> List of Tables

928	<b>2.1</b>	Sample term frequency matrix . . . . .	5
929	<b>2.2</b>	Sample tf-idf matrix . . . . .	5
930	<b>3.1</b>	Number of documents and vocabulary size for Editorials and Forums .	9
931	<b>3.2</b>	Number of documents and vocabulary size for Editorial articles and Comments . . . . .	9
933	<b>3.3</b>	Final number of topics for Editorials and Forums . . . . .	10
934	<b>4.1</b>	Labeled topics manually and with intrinsic method and . . . . .	23
935	<b>4.2</b>	Labeled topics according with intrinsic method . . . . .	23
936	<b>4.3</b>	Labeled topics with extrinsic methods and manually . . . . .	26
937	<b>4.4</b>	Label counts of non informative words . . . . .	27
938	<b>4.5</b>	Ranked similarity functions for extrinsic labeling . . . . .	27
939	<b>4.6</b>	Document topic matrix . . . . .	29
940	<b>A.1</b>	Article statistics for English editorial data . . . . .	48
941	<b>A.2</b>	Comment statistics for English editorial data . . . . .	48
942	<b>A.3</b>	Article statistics for English forum data . . . . .	49
943	<b>A.4</b>	Comment statistics for English forum data . . . . .	49
944	<b>A.5</b>	Article statistics for German editorial data . . . . .	50
945	<b>A.6</b>	Comment statistics for German editorial data . . . . .	51
946	<b>A.7</b>	Article statistics for German forum data . . . . .	52
947	<b>A.8</b>	Comment statistics for German forum data . . . . .	52