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. November 2018

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

Munich, 15. November 2018 Maria Potzner

Abstract

blablablub

Zusammenfassung

blablablub

Acknowledgement

As this thesis borders between computer science and qualitative research on consumer behaviour I would like PD Dr. Georg Groh of the Research Group for Social Computing for his input 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.

Special thanks got to my supervisor Dietrich Trautmann for his support and good ideas during the project and for the continuous reviews and feedback while I wrote this thesis.

Furthermore, I would like to thank the other team members of the SocialROM project Adnan Akhundov, Ahmed Ayad, Tim Berger, Rajat Jain, Tim Berger, Vishesh Mathur and Adrian Philipp for often tedious but every-time fruitful discussions every week.

While writing this thesis the English Writing Center of the TUM was contacted several times. I especially like to thank the fellows Rose Jacobs, Sean Rohringer, Hasan Ashraf, and Keefe Huang for reviewing my thesis.

I would like to use this opportunity to thank my parents Irina and Alexander as well as my brother Julian for their continued support during the first part of my studies. Further, I would like to thank Maria Potzner for her support while working on this project and for proof-reading this thesis.

Contents

List of acronyms

1 Introduction

This work builts up on a previous project by(zitate). when necessary this project is refereed to as Generation 1

Generation 1 (Widmer, 2018)

1.1 Research Objectives

1.2 Thesis structure

Chapter ??

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

2 Methodology

In this chapter the basic principles for the following chapters will be explained. The Section 2.1 describes how documents can be numerically represented. Section 2.2 then will introduce the three Topic Models Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) and Hierarchical Latent Dirichlet Allocation (HLDA) which are used in this thesis.

2.1 Document representation

2.1.1 Bag of Words

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

2.1.2 Tf-Idf Weighting

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

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

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

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

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

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

2.1.3 Vector space model

The representation of documents in the same vector space is known as the vector space model. This was originally introduced for Information Retrival (IR) operations like scoring documents on a query, document classification or clustering Salton et al., 1975.

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

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

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

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

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

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

2.2 Topic Modeling

Every day large amounts of information are collected and become available. The vast quantities of data make it difficult to access those information we are looking for. Therefore we need methods that help us to organize, summarize and understand large collections of data.
Topic Modeling is used to process large collections efficiently.It helps to discover hidden themes or rather topics of document collections. A topic is a multinomial distribution over all words in a corpus. Of course the probabilities over each word are different.

2.2.1 Latent Dirichlet Allocation

2.2.2 Non negative Matrix Factorization

2.2.3 Hierarchical Latent Dirichlet Allocation

3 Dataset

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

3.1 Data collection

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

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

After the selection, the chosen datasets were automatically scraped and examined for terms like *bio Lebensmittel*, *bio Landwirtschaft* for the German and terms like *organic, organic food, organic agriculture*, and *organic farming* for the English language using

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

site‘s internal search engines or Google search, which offers the option to search for sites within a domain. Nevertheless, still non relevant data like recipes, product presentations, and stock market information remained. These were kicked out by the binary Naive Bayes classifier, which was trained on 1000 random articles[3], that were labeled either as relevant or not by the Domain experts. The final collection stored in a JSON schema and the list of all sources and their percentage of relevant articles together with other descriptive statistics can be found in Appendix A.

3.2 Data processing

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

3.3 Final Datasets

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

We created two different final datasets where the frequent words, occurring over 90% in a document, and the infrequent words, occurring under 0,05%, were kicked out. The first dataset consists of editorial articles, editorial comments and forums. The final number of documents and amount of words is listed in Table 3.1. The

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

second dataset consists of editorial articles and the summarized comments from the editorials and forums. This is shown in Table 3.2. Both datasets were built for the German and English language.

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

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

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

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

3.4 Topic Generation

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

1. For every Topic Model with different topic numbers a plot was generated, see Figure 3.1. The x-axis shows the values for the most probable topic for every

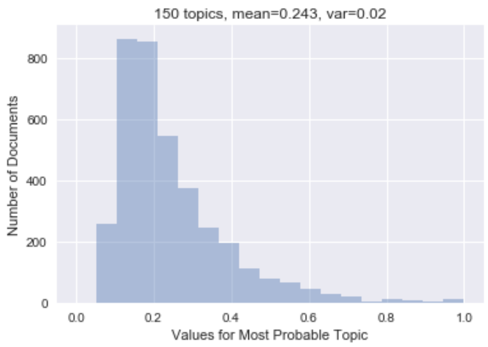


Fig. 3.1.: Count of the value of the most probable topic, summed over all topics.

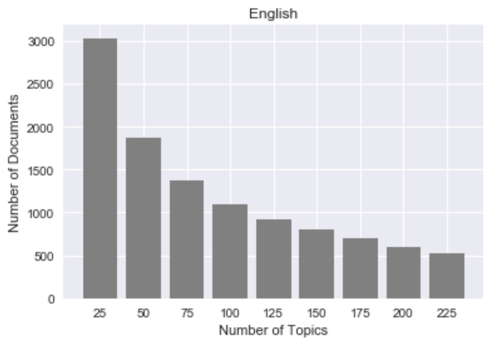


Fig. 3.2.: Number of documents the topics are expressed above the threshold

single document while the y-axis shows the counted documents where the topic occurs.

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

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

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

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

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

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

4 Experiments and Evaluation

4.1 Topic ranking

4.1.1 Related work

4.1.2 Topic Coherence

4.1.3 Theta

4.1.4 Iterrater reliability

4.2 Automatic Topic Labeling

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

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

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

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

4.2.1 Related work

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

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

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

1. construction of the semantic graph from top concepts in the given topic
2. selection and analysis of the thematic graph (subgraph form the semantic graph)
3. topic graph extraction from the thematic graph concepts
4. computation of the semantic similarity between topic and the candidate labels of the topic label graph

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

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

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

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

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

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

4.2.2 Intrinsic Topic Labeling

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

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

The candidate labels were ranked by their semantic similarity to the topic distribution θ. 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 KL divergence $D(\theta||l)$ was applied to calculate the closeness between the label and

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

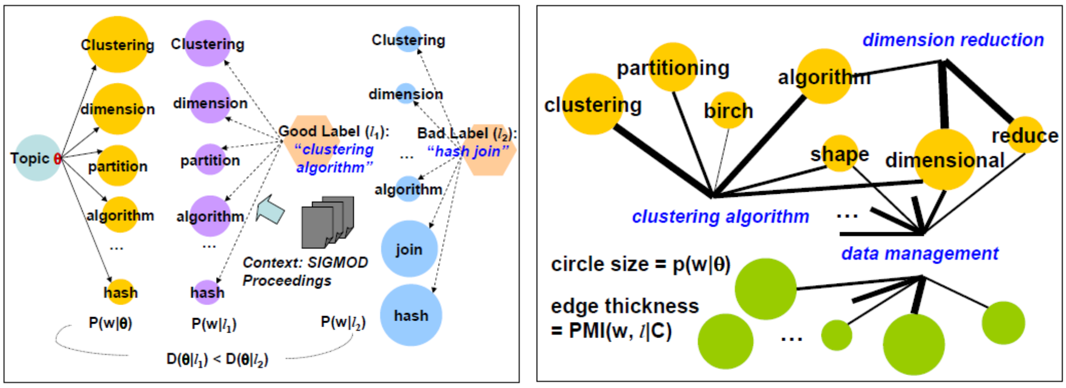


Fig. 4.1.: Relevance scoring function for ATL. Adapted from Mei et al., 2007

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

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

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

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

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

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

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

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

4.2.3 Extrinsic Labeling

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

We used the English online database *WordNet*[2], that contains 118.000 different word forms and 90.000 word senses. WordNet organizes the several types of words like nouns, verbs, adjectives and adverbs into sets of synonyms, called *synsets*. A *synonym* is a word that has the same meaning as another word. E.g *shut* is a synonym for *close*. These two words form together with possibly other words such as *fold* a synset. Additionally, a synset contains a short definition, called *gloss*, and an exemplary

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

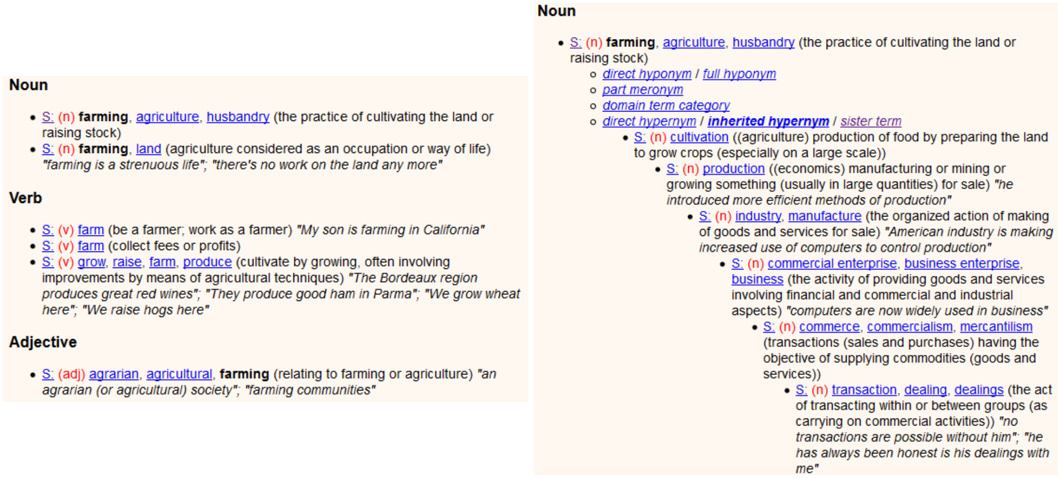


Fig. 4.2.: WordNet results for the word *farming*. Adapted from *WordNet*

sentence for each term in a synset, which describes the usage of this term. Every distinct word sense of a given word is represented as a separate synset. So the number of different meanings for a word corresponds to the number of synsets. All synsets are linked to each other according to semantic relations such as synonymy, antonymy, hyponymy, hepernymy, meronymy and troponymy. A definition of these semantic relationships can be found in Miller, 1995. In our implementation we used besides *synonymy* also *hypernymy*. If two words can be generalized by an other word, this word is called *hypernym*. E.g *animal* is a hypernym for *cat* and *dog*.

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

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

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

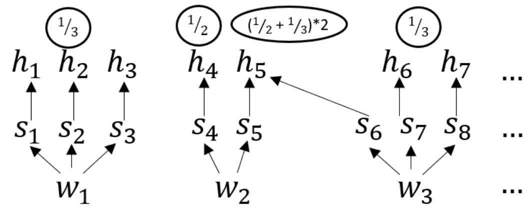


Fig. 4.3.: Scoring function for hypernyms

Preprocessing

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

Finding labels with a scoring function

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

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

Find labels with similarity functions

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

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

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

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

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

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

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

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

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

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

4.2.4 Evaluation

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

Intrinsic topic labeling

We applied the ATL in section 4.2.2 on our Dataset, which include editorials, comments and forums. In general, the ATL according to Mei et al., 2007, outputs only different labels for topics, which were generated with LDA. For the topics generated with NMF the same label was given for every topic in a topic model. The reason could be, that NMF does not have a probability distribution for θ, but although we normalized it to numbers between 0 and 1, it was not helpful. Therefore, the labels for topics generated with NMF were neglected. Further evaluations are based on English editorial articles.

Topics from Generation 1 First, we used the topics from *Generation 1*, which were generated by using collocations as described in 3.2. In Figure 4.4 the label counts for English editorial articles are shown. On the x-axis all labels are listen while the y-axis denotes the number of same labels. Considering the labels without a concrete topic assignment one can see, that they are meaningful and specific. Often, a label is a persons name e.g *Jose Andres, Rahm Emanuel, Morgan Stanly,Gloria Casas, Theresa Eisemann etc.*.

In Table 4.1 example topics are shown, which were labeled manually by domain experts and with the intrinsic approach. The intrinsic labels do not really fit to the given topic: *Rahm Emanuel* an American politician is assigned to Topic 107, which deals with environment and waste. *Hairy vetch*, a plant variety, for Topic 23. *Irritable bowel syndrome* to Topic 64 and *Safran Foer*, an American novelist, to Topic 74,

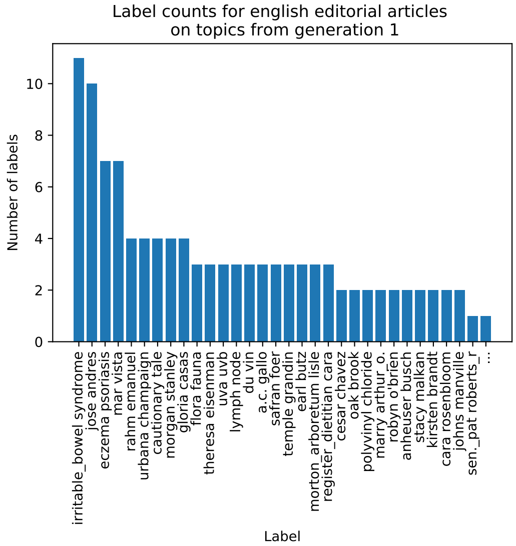


Fig. 4.4.: Label counts for topics from Generation 1 according to Mei et al., 2007.

dealing with animal husbandry. The automatic labels have nothing in common with the manual ones.

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

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

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

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

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

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

Tab. 4.2.: Labeled topics with intrinsic method

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

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

Extrinsic topic labeling

We applied the ATL in section 4.2.2 on our Dataset, using the English online database WordNet and Polyglot embeddings. The different similarity functions from WordNet, the Csf and the Polyglot embeddings were used to label our topics. A few examples are shown in Table 4.3 including the manual assigned labels to the topics, too. Some labels generated with the automatic approaches match with the manual assigned

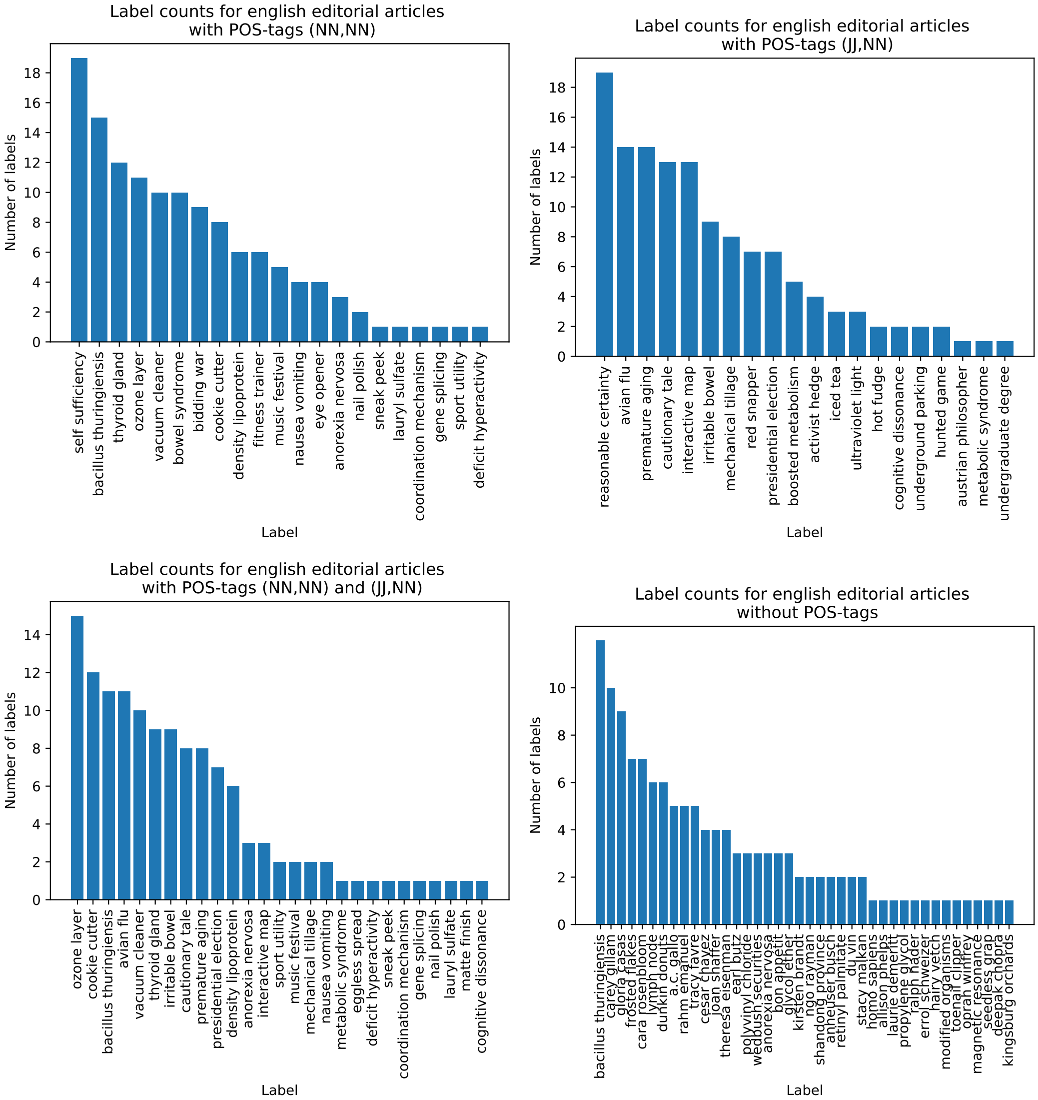


Fig. 4.5.: Label counts for topics including POS-tags with intrinsic method.

labels. This is the case for the Topics 64, 84 and 107. For the other topics, the labels heading to the right direction as the manual label: for Topic 97 *chemical* and manually *pesticide residues*, for Topic 99 *bee* and manually *beekeeping* and for Topic 109 *grocery store, mercantile establishment, marketplace* and manually *retailers* was assigned. Evaluating the automatically generated labels using different approaches, it was discovered that depending on the topics different labeling techniques output the best labels. It is not possible to tell, which approach is the best for all topics, let alone for several topic models according to the labels. Therefore, we tried to evaluate the labels generated with the extrinsic methods according to label counts. The words *entity, physical entity, object, whole, matter* and *abstraction* were chosen, because these are the most general words in the hierarchical tree of hypernyms in WordNet and do not have a high informative value. In Table ?? the number of non informative words are listed for the different similarity functions from WordNet. Based on the

	Topic 23		Topic 64	
method	grow, garden, plant, farm, vegetable, seed, year, tomato, produce, farming		milk, raw, dairy, product, cheese, cow health, drink, study, claim	
path	entity	produce	abstraction	beverage
ich	entity	produce	abstraction	produce
res	produce	produce	**dairy product**	beverage
jsn	produce	produce	produce	beverage
lin	produce	produce	beverage	beverage
plg	vegetable	vegetable	**dairy product**	abstraction
Csf	cultivate	cultivate	nakedness	farm
manual	homegrown food		**dairy product**	

	Topic 74		Topic 84	
method	meat, feed, beef, grass, eat, raise, cow, buy, make, animal		company, tea, brand, product, drink, honest, new, beverage, consumer, goldman	
path	entity	meat	**beverage**	**beverage**
ich	entity	abstraction	physical entity	substance
res	matter	meat	substance	substance
jsn	food	meat	**beverage**	**beverage**
lin	matter	meat	**beverage**	**beverage**
plg	cattle	physical entity	food	food
Csf	cattle	be	**beverage**	**beverage**
manual	animal husbandry		**beverage**	

	Topic 97		Topic 99	
method	fruit, vegetable, pesticide, produce, buy, eat, list, apple, residue, sweet		bee, honey, study, hive, year, beekeeper, plant, researcher, honeybee, colony	
path	matter	matter	organism	person
ich	matter	matter	organism	person
res	matter	matter	organism	organism
jsn	matter	matter	organism	whole
lin	produce	matter	bee	artifact
plg	fruit	entity	bee	artifact
Csf	chemical	chemical	farmer	scientist
manual	pesticide residues		beekeeping	

	Topic 107		Topic 109	
method	waste, compost, use, scrap, material, landfill, ton, environmental, throw, gas		foods, company, store, chain, market, executive, new, year, mackey, grocery	
path	material	material	grocery store	mercantile establishment
ich	abstraction	physical entity	physic entity	mercantile establishment
res	material	material	social group	mercantile establishment
jsn	abstraction	material	grocery store	mercantile establishment
lin	material	material	social group	mercantile establishment
plg	**waste**	abstraction	artifact	abstraction
Csf	convent	lowland	marketplace	marketplace
manual	**waste**		retailer	

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

sum of the non informative words per similarity function and Polyglot embeddings (plg), we ranked the different methods in Table 4.5. The top 3 are: res-similarity with preprocessing, lin-similarity with preprocessing and Polyglot embeddings.

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

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

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

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

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

Evaluated the intrinsic and extrinsic automatic topic labeling we can conclude, that the intrinsic approach generates meaningful and specific labels not really fitting to the topics and the extrinsic approach partially generates good results, which are comparable with the labels from the domain experts. Nevertheless, finding meaningful and high qualitative labels is not yet automatable. The knowledge and experience a human person has, can not be replaced by a machine.

4.3 Intern Consistency

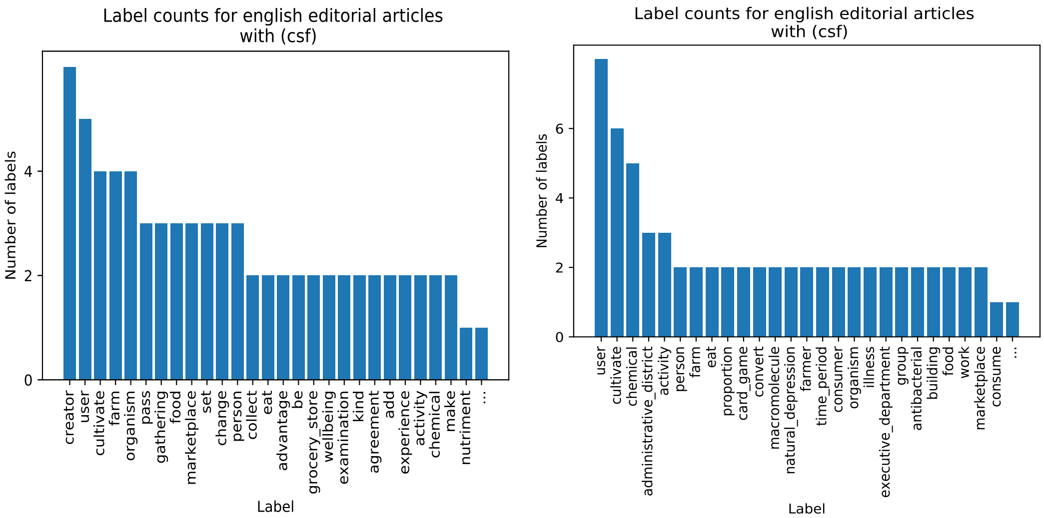


Fig. 4.6.: Label counts for topics from Generation 1 with Csf.

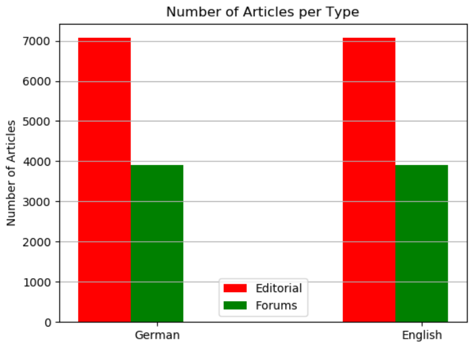
5 Future Work and Conclusion

5.1 Future work

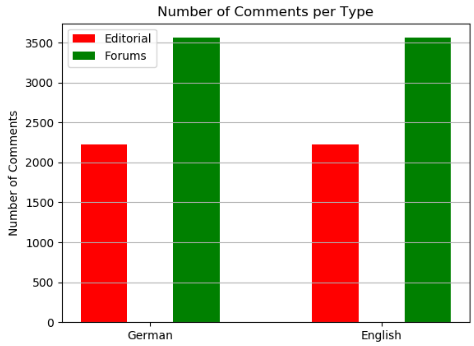
5.2 Conclusion

A

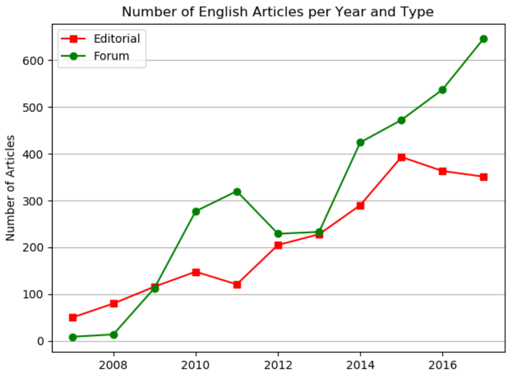
Descriptive Statistics of the Dataset



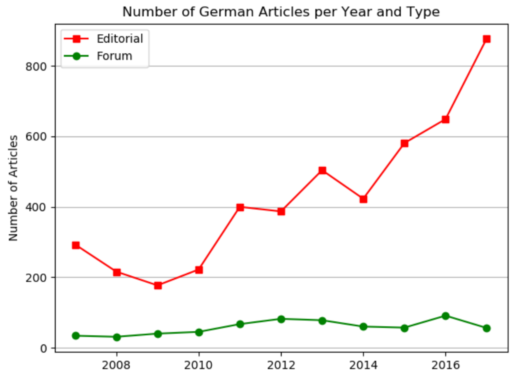
(a) Number of Articles per Type



(b) Number of Comments per Type



(c) Distribution of English articles per year



(d) Distribution of German articles per year

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

A.1 Detailed Statistics of all Sources

A.2 JSON Storage Schema

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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


Listing 1: JSON Storage Schema

Bibliography

AGOF (2018). *Nettoreichweite der Top 15 Nachrichtenseiten (ab 14 Jahre) im November 2014 in Unique Usern (in Millionen)* (cit. on p. 6).

Allahyari, Mehdi and Krys Kochut (2015). „Automatic Topic Labeling using Ontology-based Topic Models“. In: (cit. on p. 11).

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin (2016). „Automatic Labelling of Topics with Neural Embeddings“. In: 1, pp. 953–963. arXiv: 1612.05340 (cit. on pp. 12, 15).

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. 12, 15).

IVW (2018). *Verkaufte Auflage der überregionalen Tageszeitungen in Deutschland im 3 . Quartal 2018* (cit. on p. 6).

Jurafsky, Daniel and James H Martin (2009). „Speech and Language Processing“. In: *Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition* 21, pp. 0–934. arXiv: arXiv:1011.1669v3 (cit. on p. 4).

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. 12).

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. 11, 13, 15).

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. 13, 15).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to Information Retrieval*. arXiv: 05218657199780521865715 (cit. on p. 3).

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). „Automatic labeling of multinomial topic models“. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07* January 2007, p. 490 (cit. on pp. 11–14, 19, 20).

Miller, George A. (1995). „WordNet: a lexical database for English“. In: *Communications of the ACM* 38.11, pp. 39–41 (cit. on p. 16).

Salton, G, A Wong, and C S Yang (1975). „1975.A vector space model for automatic indexing.pdf“. In: 18.11 (cit. on p. 4).

Widmer, Christian (2018). „Topic Modeling for Opinion Mining“. In: (cit. on p. 2).

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

List of Figures

List of Tables