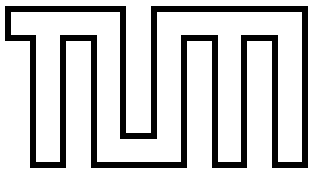


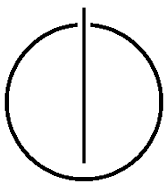
DEPARTMENT OF INFORMATICS

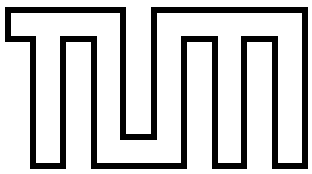
TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

Topic Model Visualization for Opinion Mining

Maria Potzner





DEPARTMENT OF INFORMATICS

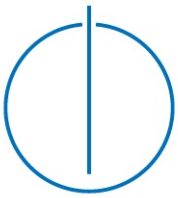
TECHNICAL UNIVERSITY OF MUNICH

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

ATL Automatic Topic Labeling 7

BoW Bag of Words 3

HLDA Hierarchical Latent Dirichlet Allocation 3

IR Information Retrival 4

KL Kullback Leibler 13

LDA Latent Dirichlet Allocation 3

NLP Natural language processing 3

NMF Non-negative Matrix Factorization 3

PMI point-wise mutual information 11

POS Part-of-speech 7

tf-idf term frequency - inverse document frequency 4

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

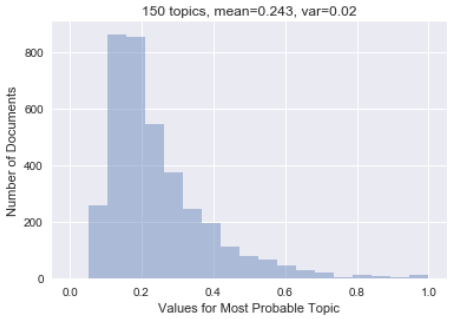


Fig. 3.1.: Count of the value of the most probable topic, summed over all topics.

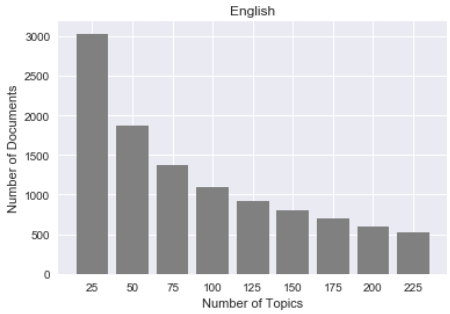


Fig. 3.2.: Number of documents the topics are expressed above the threshold

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. 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 but 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 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 every dataset the three labels 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 of 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 out of the article titles which were found in Wikipedia or Google by searching the top N words from the topics, called primary candidate labels. Afterwards, these were chunkized and n-grams were generated, which were searched in Wikipedia and Google. Theses secondary candidate labels were then filtered with the *related article conceptual overlap* (RACO), that removed all outlier labels, like stop words. Then the primary and secondary candidate labels were ranked by features like 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 key words 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. 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 for both, generated by either doc2vec(for fine-grained labels) or word2vec (for generic labels), to generate topic labels. The relevance of each title embedding was measured with the pairwise cosine similarity of the word embeddings for the top N topic words. The sum of of the relevance of doc2vec and vec2doc served as ranking for the labels. The results were evaluated the same way 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 with the topic words applying different similarity measures showing a coherent behavior with respect to the semantics of the compared concepts. The most suitable label was then selected by exploiting a set of labeling rules. This approach is applicable to any topic hierarchy summarized through a tree. The resuts were promising.

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 only based on the text collection and the extracted topics and 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 the implementation from Github[1].

In their paper *Mei et al., 2007* consider noun phrases and n-grams as candidate labels and use POS-tags to extract the label 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 semantical similarity to the topic model θ. To measure the semantic relevance between the topic and the 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 the topic distribution $p(w|\theta)$. So the KL divergence served to capture how well the label fits to our 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

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

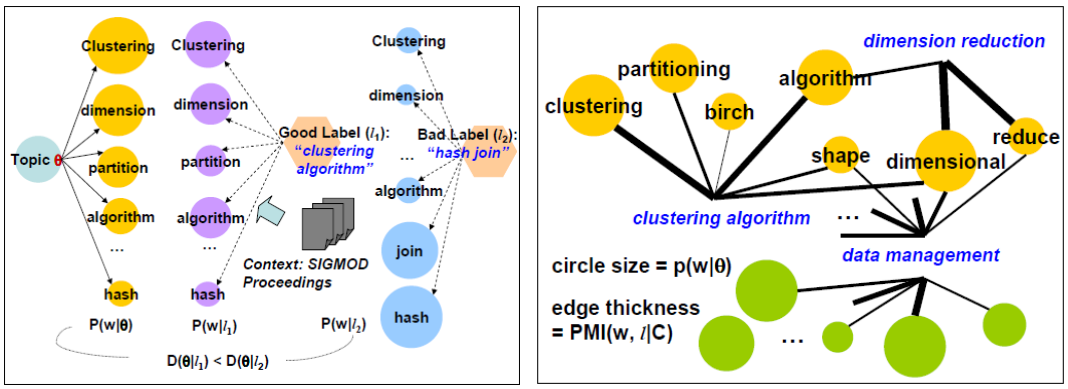


Fig. 4.1.: Relevance scoring function for ATL. Adapted from Mei et al., 2007

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 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 $p(w|l)$ the *SIGMOD Proceedings* were used as the text collection C.

Analogously, we used our datasets. On the right one can a weighted graph, where each node is a term in the topic model 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:

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

$\theta_{1,\dots,i-1,i+1,\dots}$ describes all topics except the θ_i and μ controls the discriminative power.

Evaluation

We applied the ATL according Mei et al., 2007 on the Dataset 3.3.

4.2.3 Extrinsic Labeling

[2]

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] http://www.nltk.org/howto/wordnet.html

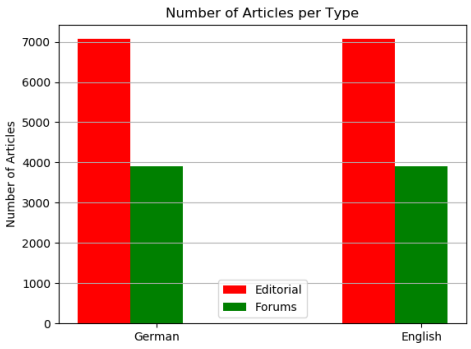
4.3 Intern Consistency

5 Future Work and Conclusion

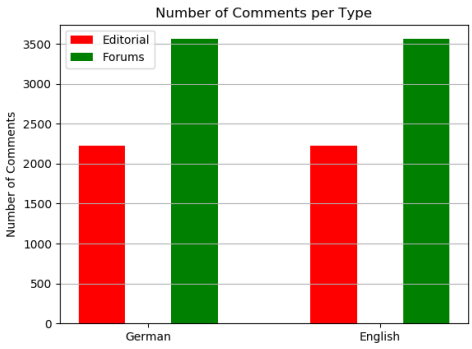
5.1 Future work

5.2 Conclusion

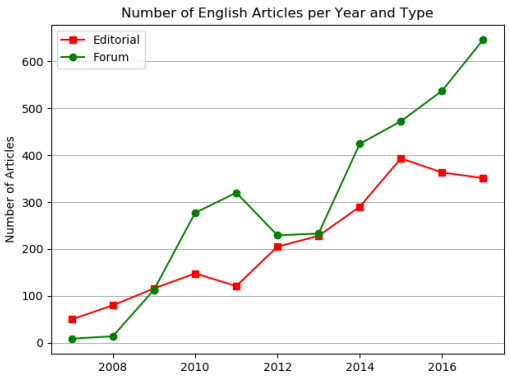
A Descriptive Statistics of the Dataset



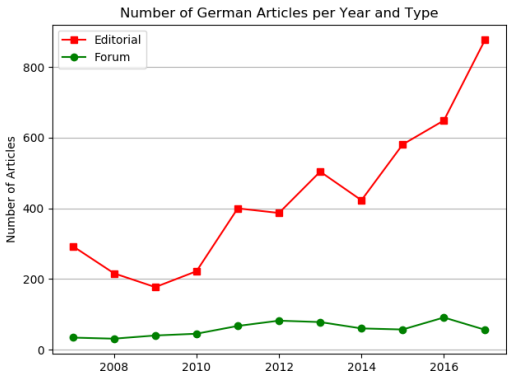
(a) Number of Articles per Type



(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

```
{
   "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 p. 12).

Carbonell, Jaime and Jade Goldstein (1998). „The use of MMR, diversity-based reranking for reordering documents and producing summaries“. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98* June, pp. 335–336.

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 p. 12).

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, 12).

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 p. 13).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze (2008). *Introduction to Information Retrieval*. arXiv: 052186571997805218657 15 (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. 11–15).

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