gesis
Leibniz Institute for
the Social Sciences

UNIVERSITÄT
KOBLENZ · LANDAU
Faculty 4: Computer Science
and Technologies

WeST
People and Knowledge Networks
Institute for Web Science

Topic Models on Biased Corpora

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Informatik

submitted by
Marcel Reif

First supervisor: JProf. Dr. Claudia Wagner
Institute for Web Science and Technologies

Second supervisor: Dr. Christoph Carl Kling
Institute for Web Science and Technologies

Koblenz, January 2018

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	☐	☐
I agree to have this thesis published on the Web.	☐	☐
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	☐	☐
The source code is available under a GNU General Public License (GPLv3).	☐	☐
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	☐	☐

.....

(Place, Date) (Signature)

Zusammenfassung

Topic Models sind ein beliebtes Werkzeug um Themen in großen Textkorpora zu identifizieren. Diese Textkorpora enthalten oft versteckte Meta-Gruppen. Das Größenverhältnis zwischen diesen Gruppen variiert meist stark. Die Präsenz dieser Gruppen wird in der Praxis oft ignoriert. Diese Masterarbeit erforscht daher ob diese Gruppen Einfluss auf ein Topic Model haben.

Um den Einfluss zu testen, wird LDA auf Samples mit unterschiedlichen Gruppengrößen trainiert. Die Samples werden von Textkorpora mit großen Gruppenunterschieden (d.h. Sprachunterschieden) und kleinen Gruppenunterschieden (d.h. Unterschiede in der politische Orientierung) generiert. Die Leistungsfähigkeit von LDA wird per "Perplexity" evaluiert.

Der Einfluss von Gruppen auf die generelle Leistungsfähigkeit von Topic Models hängt von verschiedenen Faktoren der Gruppen ab, z.B. der Vorhersagbarkeit der Sprache generell. Die Leistungsfähigkeit der Topic Models für die einzelnen Gruppen wird von der Variation der relativen Gruppengrößen beeinflusst. Allerdings ist der Effekt für alle Datensätze verschieden.

LDA kann die Gruppen intern unterscheiden, wenn die Unterschiede der Gruppen groß genug sind (z.B. Sprachunterschiede). Der Anteil der Topics, die explizit für eine Gruppe gelernt werden, ist jedoch unterproportional zu dem Anteil der Gruppe im Trainingskorpus. Dieser Effekt verstärkt sich für kleinere Minderheiten.

Abstract

Topic models are a popular tool to extract concepts of large text corpora. These text corpora tend to contain hidden meta groups. The size relation of these groups is frequently imbalanced. Their presence is often ignored when applying a topic model. Therefore, this thesis explores the influence of such imbalanced corpora on topic models.

The influence is tested by training LDA on samples with varying size relations. The samples are generated from data sets containing a large group differences i.e language difference and small group differences i.e. political orientation. The predictive performance on those imbalanced corpora is judged using perplexity.

The experiments show that the presence of groups in training corpora can influence the prediction performance of LDA. The impact varies due to various factors, including language-specific perplexity scores. The group-related prediction performance changes for groups when varying the relative group sizes. The actual change varies between data sets.

LDA is able to distinguish between different latent groups in document corpora if differences between groups are large enough, e.g. for groups with different languages. The proportion of group-specific topics is under-proportional to the share of the group in the corpus and relatively smaller for minorities.

Contents

List of Tables

List of Figures

1 Introduction

Text mining methods which help users to gain insights into larger document corpora become increasingly popular. Larger document corpora (e.g. Wikipedia database dumps) consist of millions of documents and can cover a variety of different topics (say politics or sports). While a human reader could read and understand only a limited number of documents, automated methods process large amounts of documents.

To process larger data sets, sophisticated algorithms are required. Depending on the task, the right method has to be chosen and in many cases, parameters have to be set. It is crucial to understand the behaviour of methods to make a well-informed choice on the appropriate method.

Large corpora usually provide additional information on documents: metadata, which store information on the document such as the language of a document or the source of a news article. Categorical metadata variables can be interpreted as information on *group* memberships of documents. For instance, language information might group documents into German and English documents. In practice these groups are often of unequal size e.g. the documents that are written by men exceed the amount of documents written by women in a computer science corpus.

One important class of text mining methods are topic models, which use statistical means to detect latent topics in documents. The most-popular topic models focus on the actual content of the documents and ignore any attached metadata.

The presence of groups in corpora might influence the topic detection process in documents. To illustrate, imagine you want to detect the topics of a science publication corpus. 80% of the articles are written by British scientists and only 20% of the articles are written by German scientists. Will the topic model build topics for the documents of the German scientists? How well can it predict their documents?

It is currently unknown if and to what extend such group imbalance influences the quality of topic models. While one might be tempted to predict that the minority group will have a worse prediction performance, the topic model could also detect topics for both groups proportionally to their relative share, and the group-specific model performance could be unaffected.

1.1 Research Topic

This thesis examines the following four questions:

(i) Does the presence of *groups* in corpora influence the prediction performance of topic models?

(ii) Does the prediction performance of topic models change when varying the relative *group* sizes?

(iii) Is the topic model able two distinguish between different latent groups in imbalanced corpora?

(iv) If the model can distinguish the latent groups, is the proportion of group-specific topics under-proportional to the share of the group in the corpus?

These are important questions to ask since data mining practitioners often train topic models on corpora that consist of different groups of documents and the group sizes can be imbalanced. However, this is often unknown or neglected. It is unclear to what extend the quality of topics and assignment of topics to documents is affected.

A reasonable assumption is, that the presence of groups in a training corpus influences the overall prediction performance negatively. If one of these groups is forced into a minority position, this group will additionally suffer from lower prediction performance.

This thesis should shed light on how relative group size differences affect the performance and the parameters of a topic model.

In order to answer questions (i) – (iv), multiple LDA topic models [2] are trained on imbalanced corpora. LDA or "Latent Dirichlet Allociation" is the most commonly used topic model. The imbalanced corpora are samples of a larger data set. Each sample contains two predetermined *groups*. The relative size relation of those groups is manipulated while keeping the concepts within the sample consistent.

The quality of topic models trained on these samples is evaluated using the perplexity on held-out data. Perplexity is based on the likelihood of held-out documents and explains how well the topic model can predict the test data. Therefore, perplexity is used to measure the prediction performance of LDA and answer questions (i) and (ii). The remaining questions (iii) and (iv) are evaluated by examining the predicted topic distributions over the test documents.

1.2 Thesis Structure

Section 2 describes background information of this thesis and gives an overview of the current state of research. A short introduction to topic models and an explanation of Latent Dirichlet Allocation is given. Approaches for the evaluation of topic models are presented, including the definition of perplexity. Additionally, problems related to the evaluation of semantic cohesiveness of topic models are discussed.

Section 3 illustrates the experiment. It explains in detail how the experiment is set up. This section describes how the samples are created and how the concepts are controlled when changing the relative size of groups. It provides information about the training and evaluation process which was built around the gensim [24] module.

Section 4 describes the data sets used during the experiment. It explains the thoughts behind their selection and the setup used to create them. In total, three different data sets are described: an article data set extracted from Wikipedia and two news article data sets, built using Event Registry. The Wikipedia data set contains German and English articles while the Event Registry data sets differ by using American and British news sources.

Section 5 shows the results of the experiment and answers the four main questions of this thesis.

Section 6 summarises the findings of the thesis and points out implications for the application of topic models on corpora containing groups.

2 Background and Related Work

Topic models are a common tool in natural language processing and machine learning. They are statistical models that try to capture the hidden concepts of a document corpus. They capture these concepts as so called "topics". The first part of this section will explain the general ideas behind topic models.

Afterwards, the most common topic model, LDA, is defined as it is the topic model of choice during the experiment. The model itself is explained together with the reasons why it was chosen above others.

The predictive performance of LDA will be quantified using perplexity. A definition of perplexity is given as well as a quick overview of other possible evaluation metrics. These metrics are not used during this thesis but might give some inspiration to future evaluation setups.

Lastly, a short overview about group-specific topic models is given. These topic models were explicitly created to cope with groups in corpora.

2.1 Topic Models

Topic models are a class of algorithms which exploit co-occurrences of words in documents in order to uncover hidden sets of words which explain the co-occurrence patterns and are referred to as *topics*. Probabilistic topic models explain observed documents with an underlying, hidden probabilistic model. The observed documents are assumed to be random samples from this model. In a probabilistic topic model, each document is associated with a probability distribution over a set of topics, and topics are associated with a probability distribution over the set of words. Similar documents share a similar topic distribution.

Topic models are often employed for text mining tasks, e.g. for understanding and visualising the content of large document corpora or for detecting relations between topics and other variables of interest. Additionally, topic models can be employed as a mean for dimensionality reduction (documents are mapped to a lower-dimensional topic space) [16], as input for prediction tasks, in recommender systems (e.g. for predicting semantically related tags) [14] or in information retrieval (e.g. to understand and disambiguate the topic of query terms) [33].

Example. A paper about "Data Science" might contain words from topics "Computer Science", "Statistics" and "Data Visualization". While a news article about "German Politics" might contain words from topics "Politics" and "Europe". Both

documents are described by a distribution over all topics. But the shape of the distribution will differ between the two documents. The topic distribution of a paper about "Machine Learning" might show similarities with the "Data Science" paper because it can contain words from topics "Computer Science" and "Statistics".

A single topic is the model of a hidden concept. It is a group of words that appear together in multiple documents. Regarding the previous example: words like "Regression" and "Correlation" might appear together in the "Data Science" and the "Machine Learning" papers, so the topic model decides that "Regression" and "Correlation" belong in the same topic.

The meaning of a topic is not determined by the algorithm itself. The algorithm does not know the semantic connections between words so it cannot attach a top level definition (e.g. "Statistics" or "Computer Science") to describe the topic. The top level description is usually attached by a human based on the topic models' selection of words. For "Regression" and "Correlation" one might attach the topic name "Statistics". In some cases the labelling can be hard because the topic detection only works on co-occurrence of words. So the algorithm might form a topic that is not interpretable for humans.

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation is a probabilistic topic model by Blei et al. [2, 9, 1].

In the context of text mining LDA model's variables are defined as follows. Assume there are M documents $d_1, ..., d_M$. A specific Document m is made out of N_m words labelled $w_{m,1}, ..., w_{m,N_m}$. LDA assumes that a document is a bag-of-words. Hence, only the presence of words in a document is relevant, not its position. All unique words form the vocabulary V of the model.

LDA assumes that each document contains words from different topics. Therefore, a document is a mixture of various topics. A document can be represented as a distribution θ_m over all possible topics K. The total amount of topics in LDA K need to be given as a parameter. Each single topic of index k is a multinomial distribution ϕ_k over all words in the vocabulary V.

The parameters α and β control prior beliefs of the model. Parameter α is a – typically symmetric – vector of length K and controls the prior probabilities in the topic distribution of a documents. Parameter β is a – typically symmetric – vector of length V and controls the prior probabilities in the word distribution of the topics.

A Dirichlet distribution depending on α encodes the intuition that documents only have a significant probability for a limited number of topics. A Dirichlet distribution depending on β encodes that a topic can only have a significant probability for a limited number of words.

With these prerequirements, a text corpus D consisting of M documents, each of length N_m, can then be created with the following generative process:

1. For each document m, draw a multinomial distribution over the K topics:

$$\theta_m \sim \text{Dirichlet}(\alpha), \text{ where } m \in 1, ..., M \tag{1}$$

2. For each topic k, draw a word distribution:

$$\phi_k \sim \text{Dirichlet}(\beta), \text{ where } k \in 1, ..., K \tag{2}$$

3. For all corpus word positions i, j, where i indicates the ith document and j indicates the jth position of the word in this document i:

   a) Draw the topic the word originates from:

$$z_{i,j} \sim \text{Multinomial}(\theta_i) \tag{3}$$

   b) Draw the word based on the chosen topic $z_{i,j}$:

$$w_{i,j} \sim \text{Multinomial}(\phi_{(z_{i,j})}) \tag{4}$$

2.3 Topic Model Evaluation

The evaluation step is crucial to interpret the results of the experiment. To rate the prediction performance of a trained LDA model, perplexity is used. Perplexity is a common measure used to evaluate topic models. It describes the likelihood of held-out documents regarding the trained topic model.

Unfortunately, a good perplexity value does not necessarily indicate that the detected topics are interpretable for humans. Often the rating differs to human judgement [7]. So, additional evaluation metrics are discussed.

2.3.1 Perplexity

Perplexity is a popular measure for the ability of a probabilistic model to predict new observations [17, 2]. Because models which are able to accurately predict new events are typically the desired outcome of probabilistic modelling, perplexity scores are widely used to evaluate probabilistic models. This also includes topic models.

Perplexity is an intrinsic evaluation measurement based on the log-likelihood of held-out documents. For a document d, which contains n words $w_1, ..., w_n$, and a topic model that already learned T and the topic distribution α, the log-likelihood is defined as the following formula:

$$\mathcal{L}(d) = \sum_{i=0}^{n} \ln p(w_i|T, \alpha) \tag{5}$$

To calculate the actual perplexity for a document d the log-likelihood is normalised with regards to the words n in the document d and transformed back from logarithmic space:

$$PP = exp\left(-\frac{\mathcal{L}(d)}{n}\right) \tag{6}$$

A small perplexity indicates that a topic model does well in predicting the test sample. A high perplexity indicates that topic model is "surprised" to find the word combination of the test sample. Thus, perplexity is well suited to compare different topic models that are trained on various imbalanced corpora.

Perplexity does not capture the quality of the topics though. It provides no information of the semantic context between words. Only the predictability of a document by the topic model is known.

Example. A topic can assign high probabilities to the words: "apple, car, banana, orange". A human would judge these words as semantically incoherent.

Hence, even if the perplexity on the test documents is low the topics can be not interpretable for a human. Chang et al. showed that perplexity often produces different results than human judgement [7].

2.3.2 Alternative Evaluation Metrics

Another way to evaluate topic models is rating the quality of the topics. A good topic shows semantic cohesion between the words that were assigned to it. Evaluating semantic cohesion between words is a difficult problem for an algorithm. Therefore, the quality of a topic is often evaluated by humans. A popular task given to human raters is to detect an intruder word in the set of the top-k words of a topic [7].

Unfortunately, these tests will not work for this thesis. The experiment produces hundreds of different models, each with several different topics. The evaluation by humans is not feasible for such a large number of topics. Only an automated evaluation would be reasonable.

Even though evaluating semantic cohesion between words is a difficult problem for machines, measures have been created to rate the topic quality automatically. Popular measures are the UCI coherence [21] and the UMass coherence [19].

Both measures use an external reference corpus to estimate the general word occurrence and co-occurrence probabilities. A common choice for this corpus is the English version of Wikipedia, as it contains a massive amount of documents that show the natural use of the English language. The measures differ in their approach to combine the word probabilities for a topic.

The UMass coherence uses these probabilities to compute the mean of all log-probabilities between a word given the next lower ranked word in a topic. The UCI coherence computes the mean point-wise mutual information [4] between all possible word pairs in a topic. For the calculation, a topic is often represented by its top-10 most probable words.

A study by Röder et al. [25] showed that the UMass coherence performs worse than the UCI coherence when ranking the topics by their semantic cohesion. The UCI coherence displayed a higher correlation to human judgement on all tested six

data sets. An even higher performance can be achieved when substituting the point-wise mutual information with the normalised point-wise mutual information.

Unfortunately, none of these metrics would be meaningful for the experiments conducted in this thesis. During small scale test runs of the experiment several problems occurred. One data set contains two different languages. To the best knowledge of the author, neither of the two measures have ever been applied to non-English reference corpora. Therefore, the measures were not applied to the bilingual data set because it is uncertain if the measures produce equal results with a non-English reference corpus.

The other two data sets indicated that the groups could not be clearly separated by the model, which makes the comparison of topic quality across groups impossible. Thus, the evaluation of topic quality is only feasible for corpora of English language with strictly separable topics – a condition which will hardly be fulfilled by real-world corpora. Even corpora with groups of different language do not produce a clear separation of topics, as discovered later in this thesis.

2.4 Current State of Research

Data sets that contain an imbalanced set of groups appear frequently in the real world. Examples include posts of male and female members on Twitter or contributions on Wikipedia by citizens of different countries. Typically, some of the groups show more activity than others, creating majority and minority factions. This imbalance has to be taken into account when data is modelled, which is a well-known issue from text categorisation.

There exist various approaches to solve issues caused by group imbalance; a popular solution is to increase the amount of documents by oversampling documents of underrepresented groups or to change the weighting scheme of the classifiers [15] [13].

An alternative approach by Chen et al. consists in undersampling and oversampling documents using probabilistic topic models [8]. The authors show that the classification performs better on the minority documents if the samples are generated using a topic model. But even though they have shown that topic models can be useful in terms of re-sampling, it remains unknown how relative group sizes influence topic models in general.

2.5 Topic Models for Groups in Document Corpora

There exist two kind of topic models that cope with groups in corpora: Topic models which require the explicit assignment of documents to groups and models which learn about groups in the corpus by statistical evidence. Both kind of models can be realised with groups-specific topic distributions, for instance with a hierarchy of Dirichlet distributions. A very popular alternative to hierarchical Dirichlet distributions is a hierarchy of Dirichlet processes, the Hierarchical Dirichlet Process (HDP).

In the following, HDP-based topic models are introduced and an overview of topic models for known and unknown group information is given.

2.5.1 Hierarchical Dirichlet Process

The hierarchical Dirichlet process (HDP) topic model was introduced in 2004 by Teh et al. [29, 30]. Unlike other topic models it does not need to receive the amount of topics as a parameter.

In the context of text mining the HDP model works as follows. Assume there are J documents $d_1, ..., d_J$. A document j is made out of N_j words labelled $w_{j,1}, ..., w_{j,N_j}$. Each document contains words from different topics. A document can be represented as a distribution G_j over the infinitely many topics θ_{ji}. Each topic is a multinomial distribution over a set of words.

Documents G_i can be created by repeating the following process n_j times: decide on a topic and each time choose exactly one word based on the word distribution of the chosen topic $F(\theta_{j,i})$. The topic decision process is different for each document. This can be described with following formulas:

$$\theta_{j,i} \mid G_j \sim G_j \tag{7}$$
$$w_{j,i} \mid \theta_{j,i} \sim F(\theta_{j,i}) \tag{8}$$

Even though each document should be different in terms of their topic distribution all documents should still have the same set of topics to choose from. To model this, the authors used two Dirichlet processes:

$$G_0 \mid \gamma, H \sim \text{Dirichlet}(\gamma, H) \tag{9}$$
$$G_j \mid \alpha, G_0 \sim \text{Dirichlet}(\alpha, G_0) \tag{10}$$

The documents G_j share the same base distribution G_0 which is another Dirichlet process. The concentration parameter α varies for each document. Thus each document is conditionally independent given G_0 [29]. G_0 is the global probability measure dependent on the base distribution H and the concentration parameter γ. Collapsed inference [31] and stochastic online-inference [11] allow for efficient parameter inference even for large corpora.[3].

2.5.2 Topic Models for Known Groups

A straight-forward way of modelling group-specific parameters is to model different prior distributions over topics for different groups. The three level HDP [30] is an example of such a model. It is almost identical to the standard HDP topic model, except for additional, group-specific base-measures over topics which are drawn from the global topic measure G_0 and which serve as input for the document-specific Dirichlet processes.

Another class of topic models which explicitly model group information are polylingual topic models [18, 5], which use given information about the language of documents. The model proposed by Mimno et al. [18] requires known pairs of translated documents in the corpus, while other models, like the model proposed by Boyd-Graber et al. [5], are able to detect topic translations even without such translation pairs. In their paper Boyd-Graber et al. claimed that standard LDA trained on a corpus with multiple languages e.g. German and English would be able to distinguish the two languages and assign language specific topics. This claim will be reviewed during this thesis. It will be examined if LDA can build group-assigned topics for English and German groups in a corpus.

For settings where authorship information for documents is available, author-topic models were developed by Rosen-Zvi et al. [26, 28]. It models and mixes author-specific topic distributions to explain the creation process of documents.

Another example for mixed group-specific topic distributions is the Multi Dirichlet Process (MDP) topic model by Kling [12]. It first maps documents from so-called *context-spaces* such as time or geographical location to a set of groups which then are explicitly modelled similar to the three-level-HDP. Additionally, relations between groups are modelled.

2.5.3 Topic Models for Unknown Groups

The 3-level HDP model [30] can be extended for learning about a-priori unknown groups: One could treat the group-assignment as an unknown variable which has to be learned during parameter estimation. Canini et al. presented a HDP-based model which even allows for learning more complex hierarchies, i.e. n-level HDP models [6].

3 Methodology

The goal of this thesis is to investigate the influence of imbalanced corpora on the most popular topic model, LDA. While some probabilistic models – such as clustering methods – explicitly model the presence of latent groups, standard topic models such as LDA do not model group-specific parameters. This thesis investigates if a topic model can still differentiate between the two latent groups and will rate the prediction performance in terms of perplexity.

One could hypothesise that corpora with latent groups of unequal size will influence the prediction quality of a trained topic model for minority and majority groups. This section explains the experimental setup used to investigate this hypothesis.

Figure 1 depicts the overall setup of the experiment. It can be split up into five essential steps: **(I) creation of a sample seed, (II) article sampling, (III) assignment of documents to test and training corpora, (IV) training of LDA and (V) parameter inference and model evaluation.** Each step is described in detail in the five

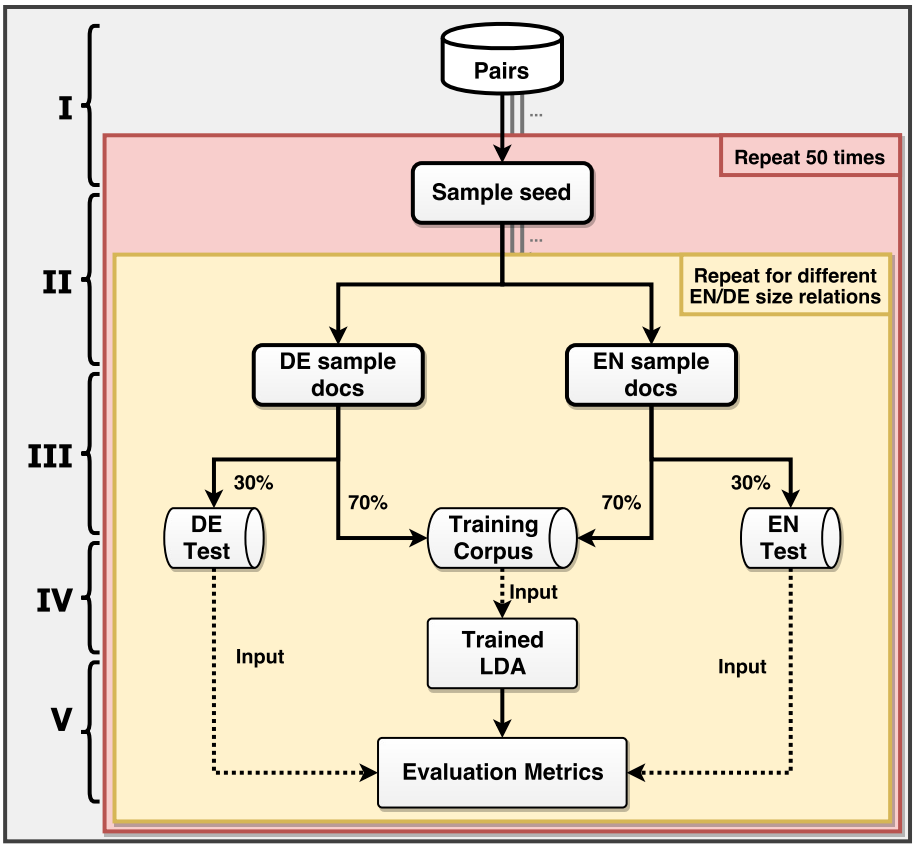


Figure 1: Outline of the experimental setup for the Wikipedia data set. Start with two groups in a data set: DE and EN. The documents in these groups have already been paired and are comparable. (I) Decide which pairs to use in a sample (sample seed). (II) Create imbalanced samples using the documents of the chosen pairs. (III) Split the documents into training and test documents by a ratio of 70/30. (IV) Train LDA using the training corpus. (V) Evaluate LDA using the test sets. Repeat steps (II)-(V) for varying relative group size relations to examine the impact of different group sizes. Repeat the whole setup 50 times to ensure that the results are not influenced by the chosen sample seed.

following sections.

The experiment is repeated using three different data sets. Their construction is explained in section 4. Regarding the experimental setup the corpora are treated almost identically. In Figure 1 and in the following sections, the Wikipedia data set is employed to explain the data set creation process. This data set's group division is more evident than the group division of the other tested data sets. The two groups that divide the Wikipedia data set are German articles (DE) and English articles (EN).

3.1 Sample Seed

During the experiment the size relation between two groups will be manipulated. At the same time, the high level concepts across these groups will be controlled. That is, during one iteration of the experiment the concepts covered in the data will stay the same. This is not possible with traditional data sets.

Therefore, during the construction of each data set, the documents have been paired (e.g. for each English Wikipedia article the German version of the article is included). This makes the root of the experiment a data set of pairs and not a data set of plain articles. Articles of a pair will be called *partners* throughout the thesis. During the construction of the data sets, it is ensured that the partners are comparable and contain the same concepts.

Based on the pairs of the initial data set a *sample seed* is created. The sample seed is a random selection of article pairs in the original data set. Sample seeds are the foundation of the samples. For each data set 50 different sample seeds are created. The variation in each sample seed ensures that the experiment tests a wide range of different topic distributions and the results do not occur by chance.

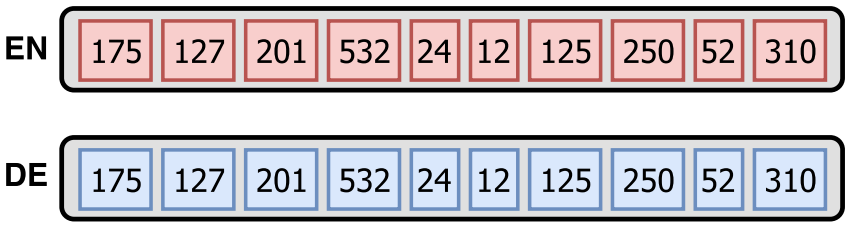


Figure 2: Example of a Sample Seed. Red articles are English articles. Blue articles are German articles. The numbers indicate the pair id.

Figure 2 shows an example of a sample seed of size 10. Each red or blue box is an article. A red box represents an English article while a blue box represents a German article. The numbers indicate the "pair id". They indicate to which pair each article belongs e.g. red article 175 is the partner article of blue article 175. The used pair ids and their order are determined randomly.

In this thesis, a sample seed size of 20,000 pairs is used for the Wikipedia corpus. The UK corpus contains 10,000 pairs and the US corpus contains 4.000 pairs.

3.2 Sampling

Articles have been paired in the initial data sets such that the partners contain the same high level concepts. In the sample seed, all articles in one group, e.g. EN 175, contain a partner article in the other group e.g. DE 175. Therefore, the concept distribution of English articles in the sample seed is equal to the concept distribution of German articles in the sample seed. In fact, the concept distribution will always stay equal as long as you select exactly as many articles as there are pairs contained in the sample seed and you ensure that there are no articles with the same pair id.

Example. If the sample seed contains 10 pairs, one can select 10 articles with different pair ids. Which group each article belongs to is irrelevant. The hidden concept distribution stays the same. This property allows building samples with different group size relations while controlling the covered concepts.

To build the actual samples a sliding threshold is introduced that indicates which articles should be included in a sample at which size relation. Figure 3 shows the concept of this threshold.

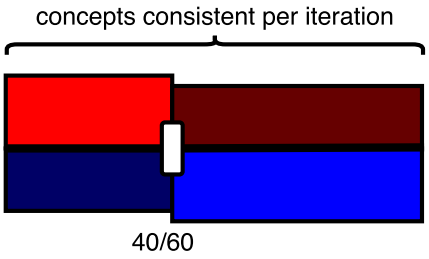
The red bar indicates English documents of the sample seed, the blue bar indicates German documents of the sample seed. The white box is the threshold. It is placed between two pair ids. Bright documents are included in a sample, darkened documents are excluded. Pair ids that occur before the threshold provide their English article while pair ids beyond the threshold will provide their German article. The sample seed contains the pair ids in a random order. Therefore, the documents themselves do not need to be selected randomly to form a random sample.

The threshold can be freely adjusted to obtain various imbalanced samples. Figure 3a shows an example of a sample that contains 40% English documents and 60% German documents. When moving the threshold further along the sample seed, a sample as depicted in Figure 3b, which contains 80% English documents and 20% German documents, can be created. During this thesis, such samples are referred to as, say, 40/60 or 80/20 samples based on size relation of the groups. The following fractions are tested: 10/90, 20/80, 30/70, 40/60, 50/50, 60/40, 70/30, 80/20 and 90/10.

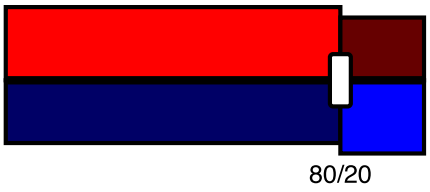
3.3 Training and Test Corpora

After sampling a sample with the desired group size relation, a training and two test corpora are created. The training corpus is used to train the topic model. The two test corpora are used to evaluate the topic model.

A part of each document is removed from the training set to use it as a test document in the test corpus. Throughout this thesis, 70% of each document are used for training and 30% of documents for testing. This ensures that the concepts that appear in the test corpus have already been seen by the topic model in the training



(a) Example of a 40% EN and 60% DE sample



(b) Example of a 80% EN and 20% DE sample

Figure 3: Creation of samples from a sample seed. Red bars describe English documents, blue bars describe German documents in the sample seed. The white box indicates the position of the threshold. Only the bright parts are contained in the sample. Darkened parts are discarded.

corpus. At the same time it prevents testing the topic model on the same documents that it was trained on.

To illustrate the process of the test and training corpus creation, Figure 4 shows the transformation of the sample seed of Figure 4 into training and test data for a 40/60 split.

Bars represent unique documents. For convenience they all share the same length in this example. In reality the size varies. Coloured bars have been selected to be part of the sample, the grey bars have been removed. The green parts of a bar show the parts that are assigned to the training corpus. The red parts of documents 175, 127, 201 and 532 are assigned to the English test corpus while the blue parts of documents 24, 12, 125, 250, 52 and 310 are used in the German test corpus. To prevent unintentionally capturing reoccurring patterns in documents, the position of the test data is chosen randomly.

3.4 Training LDA

The training corpus is used to train the topic model. The topic model used in this thesis is Latent Dirichlet Allocation (LDA). A general introduction to topic models and the definition of LDA can be found in section 2.

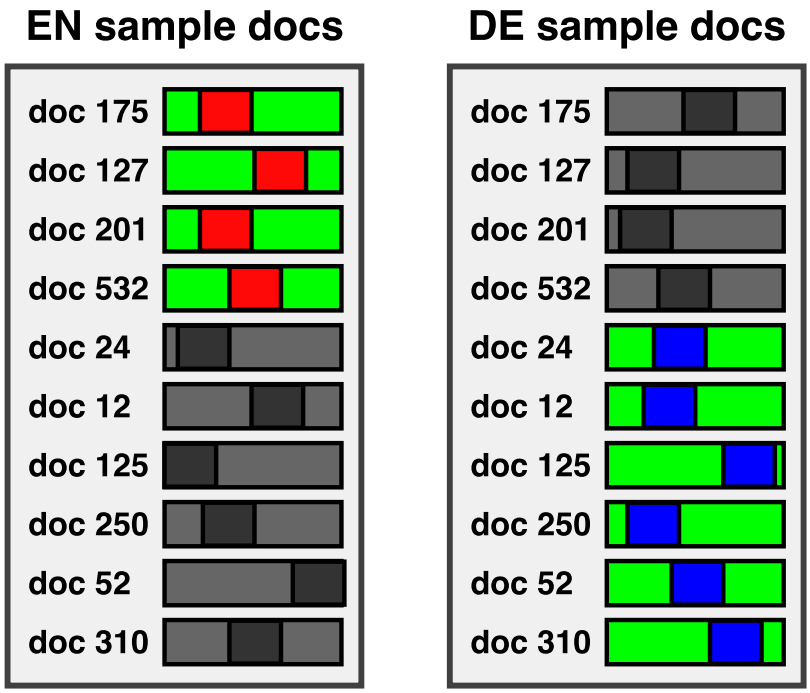


Figure 4: Example of a 40/60 sample, whose documents are split into training and test documents. Coloured bars are contained in the sample. Green parts are used during training. Red parts are English test documents. Blue parts are German test documents.

To train LDA on the test documents, the text corpus needs to be transformed into a numerical corpus. For this task the gensim module [24] is used. Gensim is a scalable and robust python module that allows efficient use of various topic models. Gensim provides an implementation of LDA, together with the methods to transform the text corpora into numerical corpora such that they can be understood by the model.

Gensim's methods are used to translate the test and training corpora. To translate text corpora into numerical corpora, a dictionary is created based of the training corpus. The dictionary maps each word in the training corpus to a number. Using the dictionary the training corpus can be transformed into a matrix. The same dictionary is used to also transform the test sets of a sample. Unique dictionaries will be computed for each sample.

The dictionaries are often very large. A large dictionary slows the training and evaluation process of a topic model significantly. So, it is reasonable to reduce the amount of words in the dictionary. It is common to remove words with low frequency in the training corpus. In this thesis, words that do not appear in a certain amount of documents are removed from the dictionary. This threshold has been adjusted to each data set's sample size. The Wikipedia sample's threshold is at 20 documents, the UK sample's threshold is at 10 documents and the US sample's threshold

is at 5 documents.

Words that appear in the test set but do not appear in the dictionary cannot be processed by the model. A dictionary might not contain a word because it was not in the training corpus or it has been removed because it does not appear in enough documents. To evaluate the impact of this loss, the vocabulary mismatch is reviewed in section 5.1.

The transformed training corpus is used to train gensim's LDA model. The implementation is based on the publication "Online Learning for Latent Dirichlet Allocation" by Hoffman et al. [10]. It uses an online variational Bayes algorithm over chunks of the training corpus to estimate the variational posterior of LDA.

LDA is a topic model that requires the user to specify the amount of topics a model should learn. To determine if the amount of topics influences the results, three different topic models are trained during the experiment. The different topic models learn 64, 128 and 256 topics.

3.5 Evaluation

The trained LDA model is used to compute the perplexity of each groups held-out test documents. A definition of perplexity can be found in section 2.3.1. The perplexity captures how likely the test documents are given the trained LDA model. It indicates how well the topic model predicts the test corpus. A low value of perplexity shows a high prediction performance. It is reasonable to expect that if a group is forced into a minority role the test documents will show a higher value of perplexity because they can be predicted worse.

Gensim's LDA model provides a method to compute the perplexity on a test corpus. The results of this function are used during the perplexity evaluations in sections 5.2 and 5.3. They answer the questions (i) if the presence of groups influences the predictive performance of topic models and (ii) if the predictive performance of the groups changes with varying the relative group sizes.

To answer the question (iii) if a topic model is able to distinguish the two latent groups, the topic predictions of the test documents are inferred from the trained LDA model. A model is able to distinguish the two groups if it is able to assign topics to a group unambiguously for both groups. A topic is assigned to a group, if the probability of belonging to a group G given a topic T is larger than 90%. That is, regarding the Wikipedia samples, for $G \in \{\text{EN}, \text{DE}\}$:

$$P(G|T) > 90\% \tag{11}$$

If the model is able to distinguish the latent groups the question (iv) if the proportion of group-specific topics is under-proportional relative to the share of the group in the training corpus can be examined.

The evaluation step completes the setup of the experiment depicted in Figure 1. The experiment is repeated 50 times for 3 different data sets, each iteration testing 9 different size relations each. Hence, the whole experiment trains and evaluates 450 different models for each data set, that is 1350 models in total.

4 Data Sets

After defining the used methods and describing the experimental setup, the data sets will be described. The data sets are the foundation of the experiment. Three different data sets are used. All data sets are constructed such that they can be separated into two groups. The separation into two groups allows the creation of samples in which the fraction of each group can be manipulated. Leading to the creation of imbalanced corpora.

In the samples, the high level concepts across the groups should be controlled. In order to make this possible, the documents in the data set are be paired. That is, a document of one group has a partner article in the other group. Both of these documents refer to the same high level concepts but use different words to do so. The choice of words to describe these concepts depends on the group of documents they belong to.

The first data set is constructed such that its groups show a high difference in the choice of words. The groups will differ by using different languages. This data set is built using a subset of Wikipedia articles in German and English language. The articles are paired using Wikipedia's language links. These links connect all different language versions of the same article. They ensure connected articles always cover the same concepts.

The other two data sets represent a more subtle scenario where the choice of words is rather similar. One data set contains US news articles, the other data set contains UK news articles. The document groups in these data sets differ in their political orientation. One group contains articles that represent a political left wing opinion while the other group contains articles that represent a political right wing opinion. The news sources and their political orientation are hand-picked.

The documents are selected using "Event Registry". "Event Registry" is a service that tracks news outlets and groups news articles to certain events. The articles are paired based on these events. Similar to the Wikipedia language links, this ensures that the articles in a pair reference the same overall concepts.

4.1 Wikipedia

Wikipedia is the largest free online encyclopedia available. It was launched in 2001 by Jimmy Wales and Larry Sanger. Wikipedia is available in nearly 300 different languages. The English Wikipedia alone counts above 5 million articles. The articles are created by users.

Wikipedia is an interesting corpus to investigate since it is one of the largest text corpora available. There are a lot of different documents covering a large variety of concepts. At the same time, it provides the tools and meta data such that documents can be divided into two groups while controlling the top level concepts.

The articles are divided based on their language. In this thesis, the Wikipedia data set has a group of English documents and a group of German documents. Other language combination might form interesting corpora as well. To find the documents

that cover the same concepts, Wikipedia's "language links" are used. Almost all articles contain "language links" which connect two articles that cover the same concepts in a different language. For example: the English article "Germany" is connected to the German article "Deutschland".

4.1.1 Data Acquisition and Formatting

Wikipedia offers regular text dumps of their articles, links and meta data. There are separate dumps for each language that Wikipedia offers.[1] The Wikipedia data set used in this thesis should contain German and English articles, hence the starting point are all articles of the English and German Wikipedia. These articles are extracted as a raw text corpus. In this thesis the latest dumps of May 2017 have been used.

Wikipedia article dumps still contain the Wikipedia specific formatting[2] i.e. HTML or Markup formatting. HTML and Markup formatting text is removed to retrieve a clean text corpus. To achieve this goal the WikiExtractor[3] by Giuseppe Attardi is used. The tool removes everything besides the section headers and actual text content of the dump and stores them together with the article title and id.

During that cleaning process several pages have been completely removed by the WikiExtractor by default. These pages were mainly forwarding pages or special pages like category or help pages.

The forwarding pages e.g. "wikipedia.org/wiki/Bombay" are empty after the WikiExtractor's format cleaning because they only include forwarding information and no actual text. The category pages, help pages, etc. – which show the URL schema "Tag:XYZ" e.g "wikipedia.org/wiki/Category:Germany" – are removed because they are often only a list of links. These links refer to pages that belong to the particular category. Help pages describe how the user should act on the Wikipedia platform and are no real articles. So their removal will not matter with respect to the experiment because only real articles should be analysed.

4.1.2 Article Pairing

During the experiment, articles need to be paired across groups. The pairs ensure that the high level concepts in a sample can be controlled while freely varying the group size relation. Paired articles are be called *partner articles* or *partners* throughout the thesis. That means, each English article will have a German partner article in the data set. Both articles will contain the same high level concepts.

To pair the articles Wikipedia's language links are used. Each Wikipedia dump contains a list of links to the different language versions of an article ("langlinks"). Wikipedia's language links do not follow a direct ID-to-ID mapping. They follow

[1] `https://dumps.wikimedia.org/enwiki/latest/`
`https://dumps.wikimedia.org/dewiki/latest/`
[2] `https://www.mediawiki.org/wiki/Help:Formatting`
[3] `https://github.com/attardi/wikiextractor`

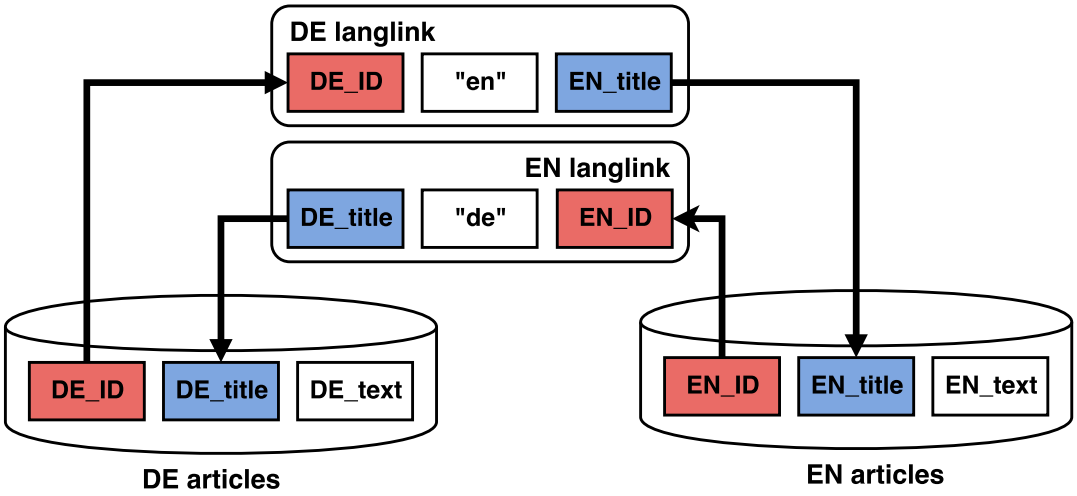


Figure 5: Wikipedia's language link schema. Blue boxes contain a language's title information. Red boxes contain a language's article ID. Based on these information the links can be resolved in the article data bases.

an ID-to-title mapping. For example, the German language links contain the language specific ID of the German article, a language identifier in ISO 639-1 standard ("en","de",...) and the title of the foreign partner article. A visualisation of this linking scheme can be found in Figure 5.

The links are not necessarily bidirectional i.e. there can be a link from a German to an English article while the link from the English to the German article is absent.

Figure 5 shows that it is possible to pair the documents based on language links. Nevertheless, the ID-to-title mapping makes it unintuitive to create article pairs. Therefore, a title look-up is created. It maps a title of an article to the corresponding English or German ID. With its help, most of the language links can be resolved such that the articles can be paired.

Some links can not be resolved. During the data acquisition and formatting step the forwarding pages, help pages and category pages have been deleted. The language links can still link to these pages. Then no text could be attached and the link is dropped. The forwarding links could have been resolved to the real pages but this can introduce new errors. It cannot be predicted to which page the forwarding page refers. Hence, the lookup could violate the ability to control the high level topics in an article pair. The following example illustrates the possible error.

Example. The English page "Square_kilometre" does not have a direct German equivalent. The language link leads to a forwarding page that links to "Quadratmeter". "Quadratmeter" is the German translation of "square meter". "Square kilometer" only has a minor subsection in the German "square meter" article.

While in this case the error would only be minor because it is generally the same concept, one cannot assume that this is always the case when there is a link between an actual article and a minor subsection.

Example. Imagine an article about a football tournament covering all teams and games. If the language link of this page forwards to a German page of "Football" where the tournament is only mentioned in a subsection or a list of tournament. Then the articles are not comparable anymore. These pages might be sharing a few of the core concepts but the concept distribution differs too much overall.

So forwarding links should not be used in the experiment and the links to forwarding pages will be dropped.

4.1.3 Link Resolution

Resolving the language links ensured that every remaining article has at least one partner. However, the partners are not always unique. Each Wikipedia page can only have one outgoing language link but it can have multiple in-links. Sometimes articles would be assigned to multiple pairs. This section explains which language link patterns can be used safely while others might introduce errors.

Figure 6 and Figure 7 summarise all possible language link patterns that can influence a pair. A node represents an article. The blue link represents the edge on which a pair is built. The direction of the edge follows the direction of Wikipedia's language link. In all patterns the pair is based on the language link from article A_1 to article B_1. In general, A_1 is called the referring article, as it is the start point of the link. B_1 is the referred article. The labels "A" and "B" represent that the articles belongs to different language groups. If "A" represents a German article, "B" represents an English article and vice versa.

The patterns depicted in Figure 6 make it impossible to build unique pairs. Figure 6a shows a scenario where the referred article B_1 does not link back to the referring article A_1. Instead it links to a different page A_2. As A_1 and A_2 are possible partners for B_1 this case cannot be resolved without further knowledge.

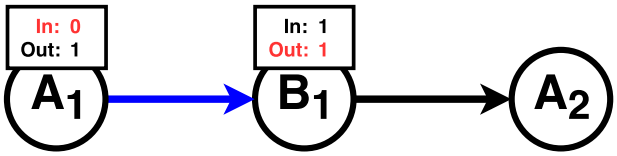
In Figure 6b the linked article B_1 is recipient of multiple links (A_1 and A_2). Again, in this case it cannot be determined which of the articles A_1 or A_2 should be the partner of B_1.

Figure 6c depicts a case where the referring article A_1 has an incoming link from an additional article B_2. This time B_1 or B_2 could be possible partners of A_1.

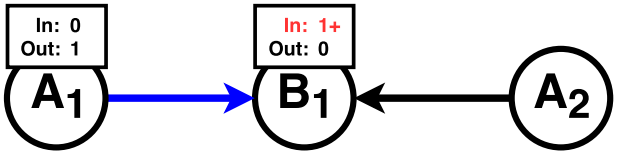
Each of these patterns suggests multiple possible pairs and will need further investigation to actually find the right pair. For this thesis, nodes involved in these patterns will be removed.

The remaining language link patterns are depicted in Figure 7 and are used to build pairs.

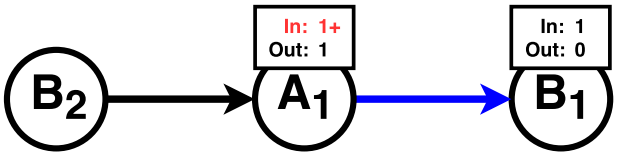
Figure 7a describes a situation where the articles A_1 and B_1 are only connected by the outgoing language link from article A_1. As long as there is no evidence against



(a) Referred article is referencing a different article

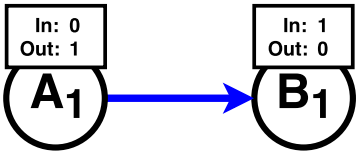


(b) Referred article is referenced by additional articles

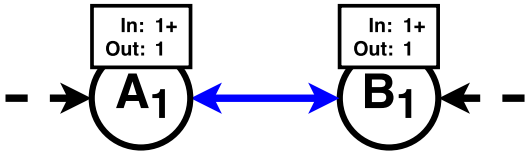


(c) Referring article is referenced by additional articles

Figure 6: Ambiguous Wikipedia link patterns. These links will not be used to build pairs. Articles contained in these link patterns are discarded.



(a) Unique link without any additional referring article



(b) Bidirectional link (disregarding additional articles that might reference an article of this pair)

Figure 7: Unambiguous Wikipedia link patterns. These links will be used to build pairs in the Wikipedia dataset.

this connection (e.g. a case from Figure 6) both articles are paired.

The final case, depicted in Figure 7b, is the optimal case. Both articles are refer-

encing each other. If this bidirectional connection is present the articles A_1 and B_1 are paired regardless of other articles linking to these pages. Even if there are other articles linking to the pair from the outside it is more likely that the bidirectional connection captures the actual pair.

4.1.4 Article Pair Filtering

Even if the articles can be paired, not all article pairs are suitable for the experiment. To ensure that the articles contain enough information and that articles in a pair are comparable in terms of their content, two restrictions will be applied:

1. If one or both articles of a pair contain less than 50 words the pair is removed
2. If one article is more than twice as long as its assigned partner the pair is removed

The first restriction forces a minimal amount of information in each article. If an article is too short it only contains few information. This is especially meaningful when splitting the documents into training and test document. The test document contains 30% of the original document. Hence, a short article might not contain enough information to build a reasonable test document. Therefore, a threshold of 50 words is introduced. If a pair contains at least one article that is shorter than 50 words it will be removed. Both groups, German and English articles, contain such documents. In total 262,415 pairs fall below this threshold and are discarded.

The second restriction applies to the length difference between the articles of a pair. If the length of article partners differs too much then the concepts covered in both articles might differ. Varying concepts between partnered articles violate the pre-requirement which is necessary to control the concepts in a sample. Big differences between the concepts can cause flawed results.

Example. Imagine a pair where the German article has 5000 words and the English article has 100 words. Both articles will share the core concepts but the longer article will go much more into detail and contain additional concepts. Consequently, the articles are not comparable.

To prevent this phenomenon, the size difference between partnered articles is restricted. If one article has twice the size of its partner article (or longer), the pair will be removed. In the whole data set 557,875 pairs do not meet this requirement. In total 651,092 pairs do not meet one or both restrictions and are removed.

The documents in these pairs are stemmed and stop words are removed. A description of this process can be found in section 4.3. The final data set is described in section 4.4.1.

4.2 Event Registry

Event Registry is a service that tracks news articles. The company tracks more than 100.000 news sources in 15 languages. Each new article that is released by one of these news sources is analysed. During the analysis, Event Registry detects article groups that report the same event using machine learning.

These event-assigned article groups allow building article pairs based on these events. The assumption is, that articles that report the same event refer to the same concepts. Therefore, the event assignment can be treated like a language link in the Wikipedia data set.

The group assignment in the Event Registry data sets is done based on the political orientation of the news sources. The data set differentiates between right-wing and left-wing sources. The first data set will focus on American sources and will be referred to as the US data set. The second data set will focus on British sources and will be referred to as the UK data set. The creation process of both data sets only differs in the initial selection of news sources. Thus, this section describes the creation of both data sets.

4.2.1 Source Selection

Event Registry tracks articles from multiple languages. In this thesis only articles written in English language are used. American and British articles are treated separately which leads to the creation of two different data sets, the US data set and the UK data set.

Each data set contains groups that differ in their political orientation. A data set contains articles with a right-wing orientation and articles with a left-wing orientation. The political orientation of a single article cannot be determined per se. Nevertheless, the general political orientation of the news sources that publish the articles can be assessed. The assignment of political orientation is based on two surveys that tried to determine the political orientation of several news sources.

The first survey was conducted by YouGov [27]. YouGov is an international market research and data analytics company. It's survey will be used as the foundation of the UK data set. The survey does have no influence on the US data set.

In 2017 YouGov conducted a survey in which they asked a sample of 2040 citizens of the UK how they would rate the political orientation of 8 mainstream newspapers. The results (excluding between 39-49% of respondents who answered with "don't know") are displayed in Figure 8.

Figure 8 shows a left-wing orientation for "The Guardian" and "The Mirror". Therefore, they will be categorised as left-wing in the data set. "The Times", "The Telegraph", "The Sun", "The Daily Express" and "The Daily Mail" show a right-wing orientation and will be categorised as such in the data set. "The Independent" is a corner case. It is not as left as the other two sources but based on its difference to the "right-wing" articles it will still be considered "left-wing". This categorisation will be extended by the second survey.

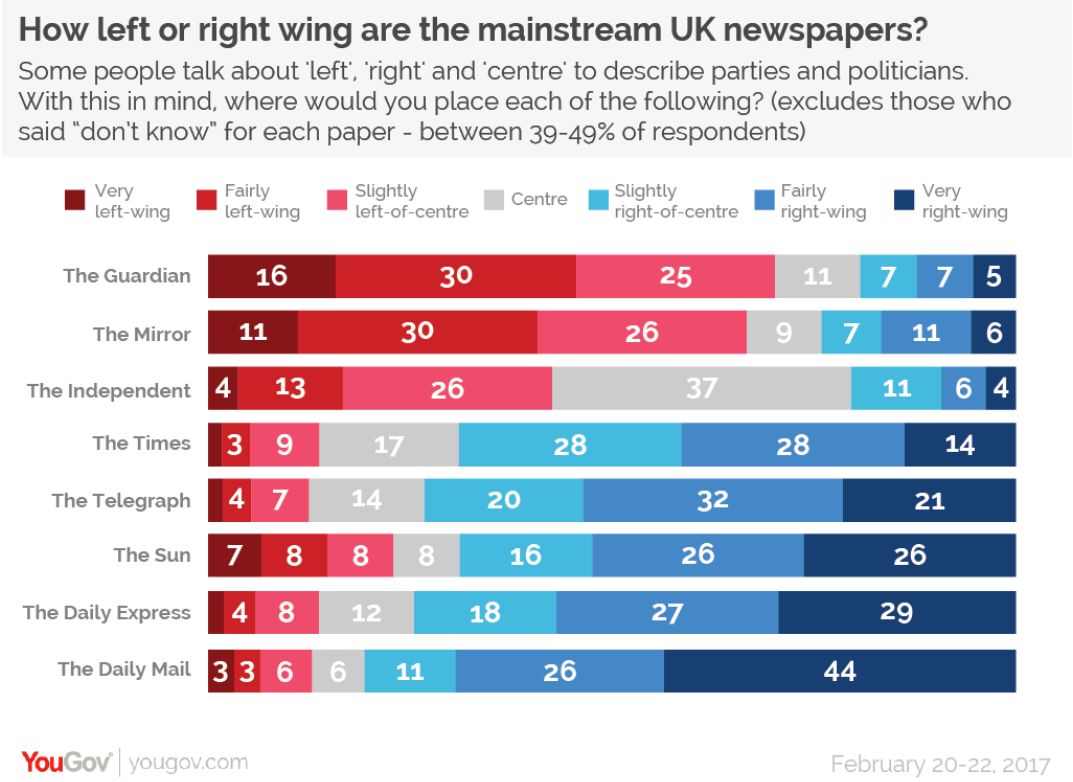


Figure 8: Political Orientation based on YouGov survey [27]. This survey is the foundation of the UK news source categorisation. "The Guardian", "The Mirror" and "The Independent" are categorised as "left"; the remaining news sources are categorised as "right".

The second survey was conducted by the Pew Research Center [20]. The Pew Research Center is an American nonprofit, nonpartisan "fact tank". They conduct data-driven social science research. While extending the UK data set, this survey will serve as the foundation of the US data set. It is used to extend the UK data set because it contains sources that are commonly perceived as British news sources.

In 2014 the Pew Research Center asked a sample of 2901 web respondents a series of 10 political values questions. They analysed the audience of 36 news sources and created an ideological profile for each audience. Figure 9 shows the average ideological orientation of each audience on a scale comparing it to the ideological placement of the average respondent.

The group of news sources in the range of "MSNBC" to "Wall Street Journal" have been considered as too central and are not used in this thesis. All sources whose ideological orientation is left of this group will be considered "left-wing", all sources that lay right of this group will be considered "right-wing".

Even though this survey was targeted for American audiences "BBC" and "The

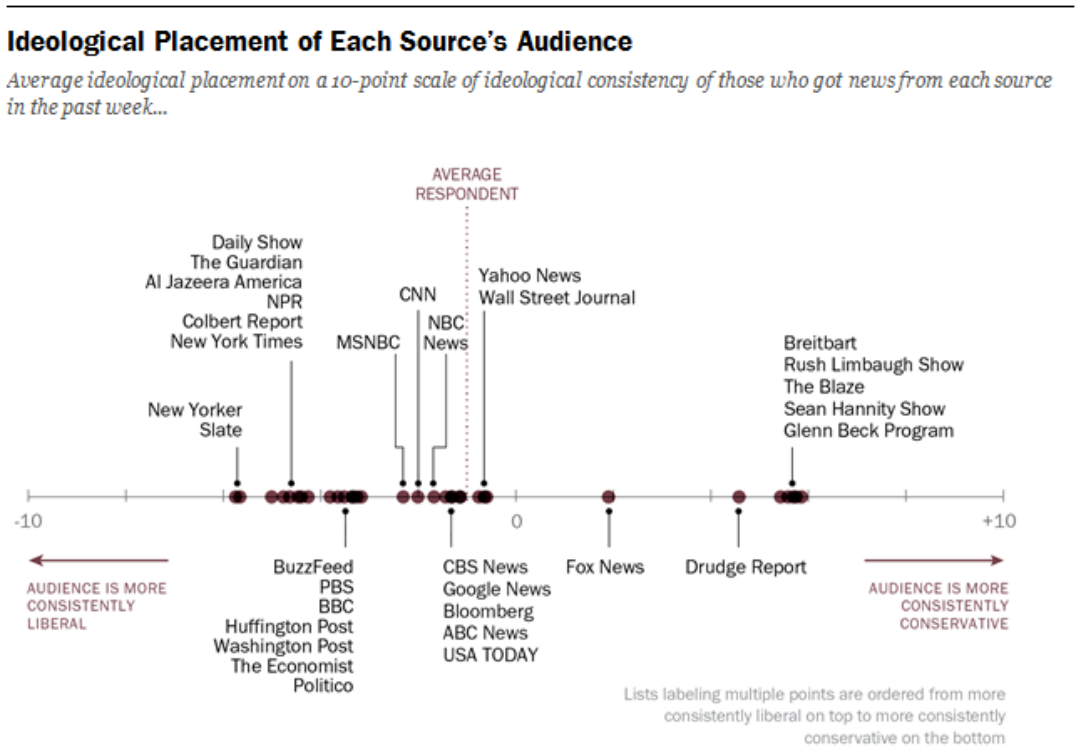


Figure 9: Ideological Profile of Each Source's Audience based on Survey of Pew Research Center [20]. This survey builds is the foundation of the US news source categorisation. Sources assigned left from "MSNBC" are categorised as "left"; sources assigned right from "Wall Street Journal" are categorised as "right".

Guardian" are primarily British news sources and will thus be used in the UK data set. They are not used in the US data set. "The Huffington Post" has two different web appearances. One tailored for the UK market and another appearance for the American market. The respective versions are used as news sources for the according data set. This leads to the final assignment of news sources for the UK data set depicted in table 1.

The remaining news sources are used to create the American data set as far as they are available at Event Registry. Using just these sources, there is a heavy overhead of left sources. To even the count between left-wing and right-wing sources additional right sources are added. These sources are hand-picked.

Table 1: Categorisation of UK online news into left and right political orientation. Articles from a left and a right news source on the same subject are paired in the experiment to construct the UK data set.

LEFT	RIGHT
theguardian.com	thetimes.co.uk
independent.co.uk	telegraph.co.uk
bbc.co.uk	thesun.co.uk
bbc.com	dailymail.co.uk
huffingtonpost.co.uk	express.co.uk
mirror.co.uk	

The following news sources: "New York Post", "World News Daily", "Newsmax" and "Townhall". To assure that these source have a right bias, the news outlets were checked on "mediabiasfactcheck.com" regarding their political tendencies. The website "mediabiasfactcheck.com" aims to define the credibility and political orientation of news outlets. They are funded by advertising and individual donations. The credibility of this website is not comparable to official survey companies like YouGov or the Pew Research Center but their judgement matched the results of the two surveys. The final list of news sources for the US data set is shown in table 2.

The source lists 1 and 2 are used to create article pairs for their respective data set. A pair will contain an article from a left source and a right source. The Assumption is, that that left-wing articles differ from right-wing articles and that the "strength" of the political orientation within each group is equally pronounced. With this assumption all pairs can be treated equally e.g. an article pair from "The Guardian" and "The Times" will be treated the same as an article pair from "BBC" and "Daily Mail".

Table 2: Categorisation of US online news into left and right political orientation. Articles from a left and a right news source on the same subject are paired in the experiment to construct the US data set.

LEFT	RIGHT
newyorker.com	breitbart.com
slate.com	foxnews.com
nytimes.com	rushlimbaugh.com
npr.org	theblaze.com
pbs.org	nypost.com
washingtonpost.com	wnd.com
buzzfeed.com	newsmax.com
politico.com	townhall.com
huffingtonpost.com	

4.2.2 Data Acquisition

With the news source lists the data sets can each be divided into two groups based on their political orientation. This section explains how these sources' news articles are acquired to build the UK and US data set.

The articles are requested from Event Registry. Event Registry offers an API that lets the user search for events and articles respectively. For this thesis, the Event Registry team provided the author with a 5000 request token license. Using this license data in a time frame of about two and a half month , starting at 10.07.2017 until 25.9.2017, was captured. The goal was the selection of events that contain at least one article from a left source and an article from right source.

The author did not find a way to extract the necessary data with a single request. You can request events that meet the condition that they contain left and right sources but you cannot retrieve articles associated with this event in the same step. To solve this issue the request has been divided into two steps:

1. Finding events that contain at least one left source and one right source (disregarding US and UK separation at this point)
2. Retrieving the articles of the according events

The first step can be implemented in a single request. The request retrieves all event ids that contain articles from a left and right source in a time frame of one week. Each week containes between 1500-2500 events. Each of these events has to be called separately to retrieve their articles. The maximum amount of articles that can be retrieved in one request is 200.

Since the total amount of available requests was limited for this thesis, the main goal of the data extration was, to minimise the amount of requests necessary for each event while still extracting the articles needed to build a pair. This led to the following extraction process:

1. Pick an event id from the events that contain a left and right article
2. Fetch 200 articles using the event id
3. If the fetched articles contain a left and right article pair, you are done
4. Else: Go to 2 (or mark as unsolvable if request threshold is reached)

The process tries to keep the amount of requests per event as small as possible by stopping early when at least one article pair can be build. This is a reasonable step to take as only a single article pair per event necessary. If it is possible to create additional pairs, these pairs are created but they are not mandatory for the experiment. In general, this process aims to create article pairs that cover a lot of events. It does not try to retrieve all articles of an event.

To reduce the amount of requests per event even further, the order in which the articles are received is adjusted. Articles are received in order of their popularity.

Event Registry offers four different sorting methods that decide in which order the articles will be returned: “Date”, “Relevance”, “Shares on Social Media” and “Story Centrality”. For this thesis “Shares on Social Media” is used. This sorting method increases the chances of getting a left and a right article in the first requested 200 articles due to the fact that all chosen news sources are rather popular or controversial.

This choice introduces some degree of selection bias as it alters the order in which the news articles received. Popular news sources are covered more frequently because unpopular news sources are sometimes not included in the first 200 received articles. However, based on the previous assumption that articles from news sources with the same political orientation are equal in their “strength” of ideological faith, the sources are interchangeable and the influence can be disregarded. Nevertheless, if there is no request limit, receiving the articles sorted by their “Date” is preferred as it will contain less bias.

After the articles of an event have been retrieved and at least one pair of left and right articles has been found, the articles can be paired. If exactly one left and one right article is found, they are paired directly and the next event is processed. If multiple pairs are possible, then the left and right articles will be paired randomly.

Event Registry sometimes contains articles that are assigned to multiple events. For sampling only single occurrences of an article are desired. So if articles appear in multiple pairs, one pair is picked randomly while the remaining pairs are discarded.

4.2.3 Article Pair Filtering

The article pairs of the Event Registry data sets are treated with the same restrictions that were introduced in section 4.1.4. An article has to have a minimum length of 50 words to ensure that an article contains enough information to be split into a training and test document. Furthermore, in a pair of articles one article can only be at maximum two times as large as its partner. If the length difference between articles in a pair is too large, the contained information and the high level concepts will vary such that the articles are not comparable anymore. Pairs that do not meet these requirements will be dropped.

Only 69 article pairs in the UK data set contain an article with less than 50 words. The US data set contains only one such article. In the UK data set 11.827 pairs do not meet the length difference restriction. In the US data set 3.975 pairs do not fulfil it. In total, the restrictions leads to the removal of 11.857 UK pairs and 3.975 US pairs.

The documents in the remaining pairs are stemmed and stop words are removed. The description of this process will follow in section 4.3. The final data sets are described in section 4.4.2 and 4.4.3.

4.3 Stemming and Stopword Removal

After building the three data sets all articles still contain a lot of words that do not contain any information about the content of the article e.g. “the”,“and”, etc. These

words are called stop words and appear in almost every document. Due to their high occurrence frequency they often influence the topics that a topic model builds. In this case, the topic model builds topics that are based on syntactic conventions and not on semantic connection. To prevent this effect stop words are removed from the corpora.

Before removing words, the initial structure of the document is saved. This information is important to correctly split the documents into a test and training document during the experiment (section 3.3). The test split will take 30% of each documents at a random position. Therefore, the the initial structure of the document should stay intact.

To save the structure, the original length of each document is stored and a position identifier is added to each word e.g. "I love data" will be transformed to [("i",0), ("love",1), ("data",2)].

To remove the stop words predefined stop lists are used [16]. Two different stop lists are used, one for English and one for German documents. The foundation of these stop lists are the stop lists provided by the natural language toolkit (nltk) [34]. The nltk module is a powerful module to process text corpora developed by Steven Bird and Edward Loper. When using only the nltk lists there were words remaining that could be considered stop words so these stop lists have been extended with additional, external German [4] and English [5] stop lists.

Another issue is the amount of different words used in the data sets. The topic model will need a vocabulary that matches every unique word to a number. So the vocabulary and the according test and training corpora will be extremely large and the subsequent computation is very slow. Therefore, we stem all words in the data set.

Stemming reduces words to a base form [16]. The base form does not necessary match the linguistic root of a word. For example, a stemmer can match the words "families" and "family" to the base form "famili". An aggressive stemmer e.g. the Lancaster stemmer [22] can sometimes reduce the words too much such that they are not easily interpretable for humans. In this thesis nltk's implementation of the Snowball stemmer developed by Martin Porter [23] is used. It can be applied on both German and English corpora and keeps word stems interpretable.

4.4 Data Set Description

This section will describe the final corpora of the Wikipedia data set, the UK data set and the US data set. This data is used during the experiment to investigate the four lead questions (presented in section 1.1). The pairs in all corpora have been filtered such that the partner articles are comparable. Stopwords have been removed and all words have been stemmed. A brief overview of the datasets can be found in Table 3.

[4] Additional German stop word list: `https://github.com/solariz/german_stopwords/blob/master/german_stopwords_full.txt`

[5] Additional English stop word list: `http://xpo6.com/list-of-english-stop-words/`

Table 3: Overview of data sets used in the experiments. Documents of each group were paired with documents of the other group on the same subject.

Name	#Pairs	Description	Groups
Wikipedia	385,306	Wikipedia articles	German and English
UK	18,170	UK news articles	Political orientation (left/right)
US	4,721	US news articles	Political orientation (left/right)

4.4.1 Wikipedia Data Set

After building article pairs the original corpus contained 1.036.398 million pairs. A lot of them did not meet the established length requirements such that the final corpus only contains 385.306 pairs of English and German Wikipedia articles. The English documents have a mean length of 522 words and the German documents have a mean length of 476 words. The cumulative document length distribution is depicted in Figure 10. Both length distributions are similar. Hence, there was no need to balance the document lengths across groups.

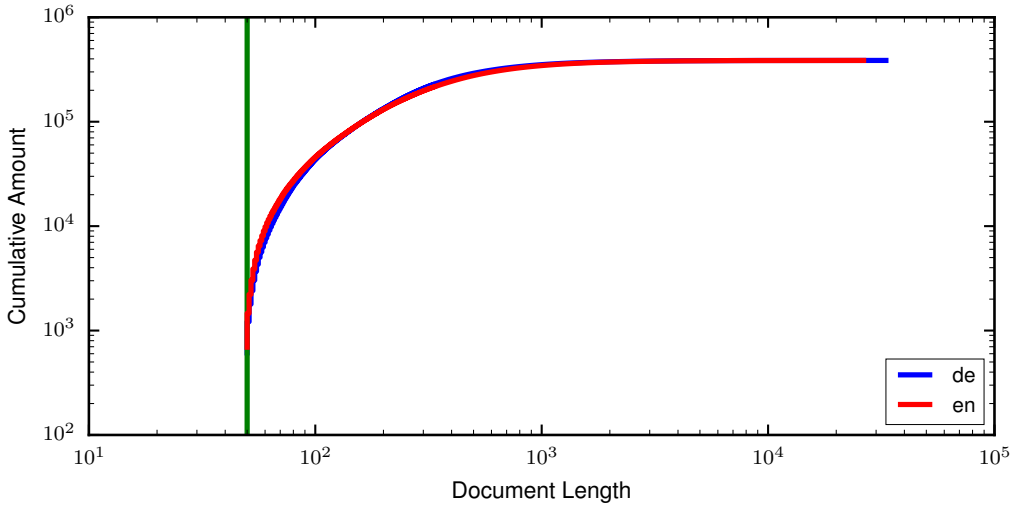


Figure 10: Article length distribution of English and German Wikipedia after length reduction. The red graph describes English articles (en) and the blue graph describes German articles (de). The green line shows the minimum length of 50 words we required for all articles. A lot for short articles up until 1000 words for both groups. Longer articles are rare. Both article length distributions are very similar.

Looking at the pairs in a corpus, an English document is longer than his German counterpart in 56.1% of the cases. The German article is longer in 43.4% of the pairs.

Only in 0.4% of all cases the German and English article are of equal size. In the average pair, the English article is 6% longer than the German partner.

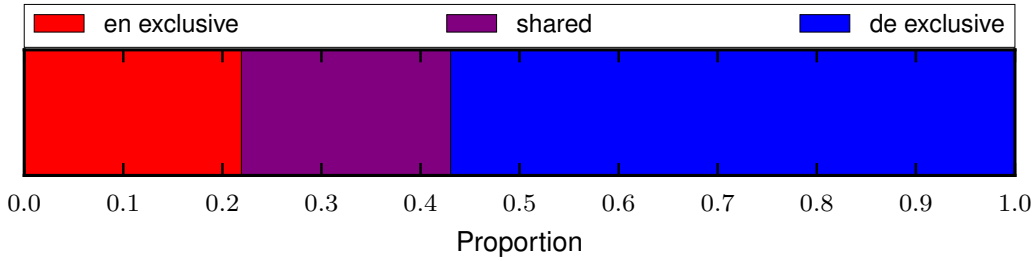


Figure 11: Distribution of unique and shared words in the Wikipedia corpus. The red bar shows words that appear only in the English group, the blue bar shows the words that only appear in the German group. The purple bar shows the words that appear in both groups. The German exclusive vocabulary is significantly larger than than the English exclusive vocabulary. Indicating that German documents will be harder to predict for a topic model.

The stemmed and stop word removed Wikipedia data set contains 2.725.922 unique words. Their distribution in German and English articles is shown in Figure 11. 1.173.116 of these words are used in the English articles. 2.128.509 of these words are used in the German articles. Both groups share 575.703 of all words, which is about 21% of the whole corpus. The vocabulary of both languages is clearly separated. On average a pair shared around 40 words.

The German vocabulary is significantly larger than the English vocabulary. An explanation for this are for example compound words, which are frequent in the German language and cannot be reduced to the base form by a stemmer. It is reasonable to expect that the increased vocabulary makes it harder for a topic model to predict the German documents during the experiment.

Figure 12 describes the word frequencies of English and German words. The words have been ranked based on their frequency beforehand. Both graphs resemble a power law function. This phenomenon is typical for natural-language documents and known as Zipf's law [17]. The highly ranked English words are more frequent than the highly ranked German words. With increasing rank the word frequency decreases faster for the English documents. Approximately, beyond rank 10.000 the rare German words are more frequent than the English words.

Thus, German documents contain more unique words in general and the rare words show a higher frequency than the English rare words. The increased frequency of rare words can have an influence on the created topics, as the topic model needs to group more words. It is expected that these additional challenges make German articles harder to predict for a topic model.

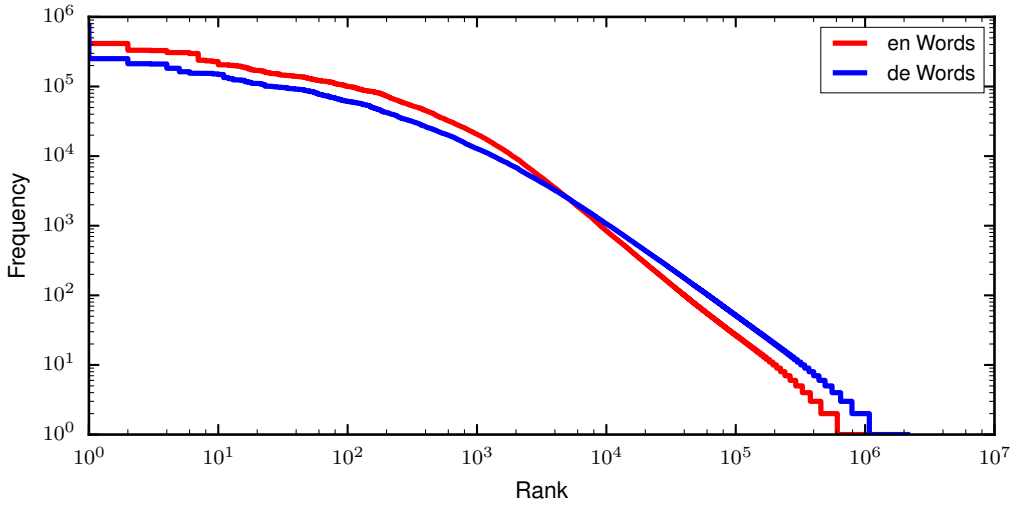


Figure 12: Word frequency in English and German Wikipedia articles sorted by word rank. The red graph represents the English words. The blue graph represents the German words. English high rank words are more frequent than the German high rank words. Words ranked beyond rank 10.000 are more frequent in German.

4.4.2 Event Registry UK Data Set

The UK data set contains 18.170 pairs divided in articles of left and right political orientation. The left and right articles were build from different combinations of left and right sources. The mean document length for both political orientations are similar. The documents of the left sources have an average length of 476 words, while the documents of the right sources have an average length of 480 words.

Figure 13 shows the distribution of sources over all article pairs. The most used left source was "The Mirror". The most common right source was "The Daily Mail". Both make up the most frequent source pair of all article pairs. Combinations of less common sources are very rare in general e.g. "The Telegraph" and "The Independent". The frequency difference will not matter for the experiment because the assumption is that all articles inside of the left or right source list are interchangeable.

The cumulative length distribution is depicted in Figure 14. The majority of documents has a length below 1000 words. Documents longer than 1000 words are rare. The right articles show an irregularity at about 125 to 150 words in a document. These documents are surprisingly frequent. It is probable that this is caused by a length threshold of one or more of the news sources. As it did not interfere with the experiment, it was not investigated further.

In 51.7% of all pairs a right article is longer than a left one and in 48.1% the left article is longer. Only in 0.2% of all cases, both articles are of equal size. In the average pair, the right article is longer by about 2%.

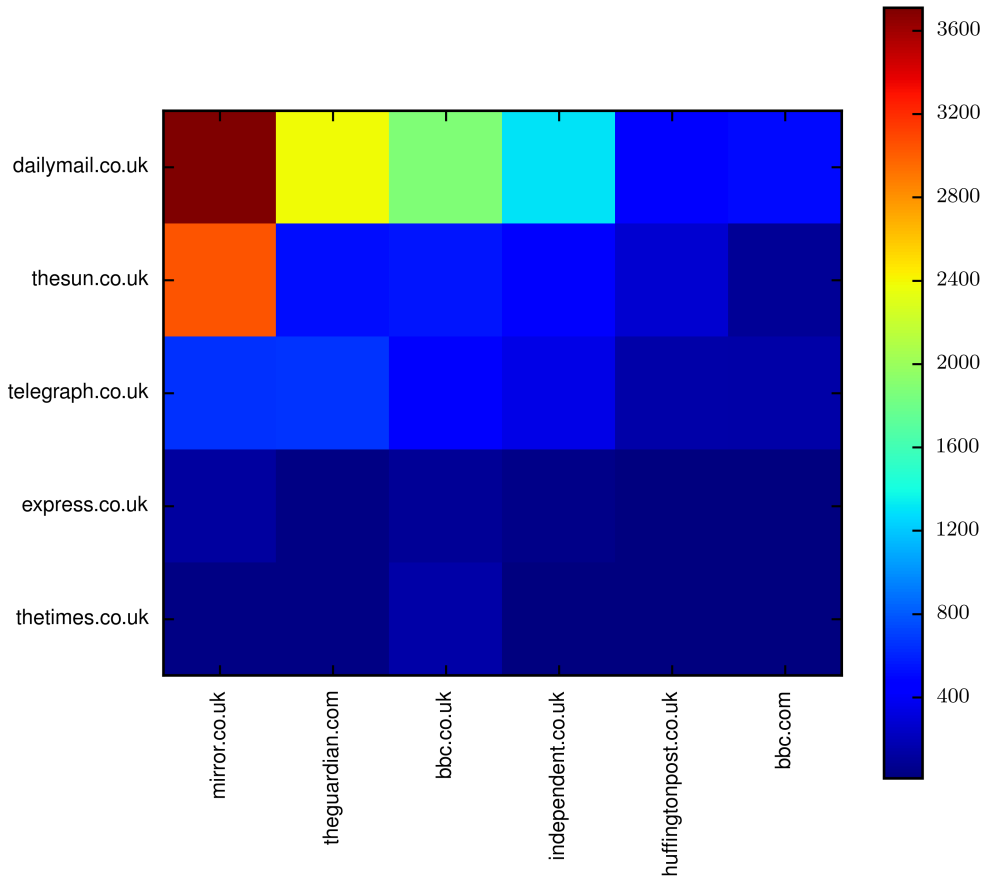


Figure 13: Source distribution of left and right UK articles. Red shows a high frequency of a pair. Blue indicates low frequency. "The Daily Mail" and "The Mirror" are the most common pair.

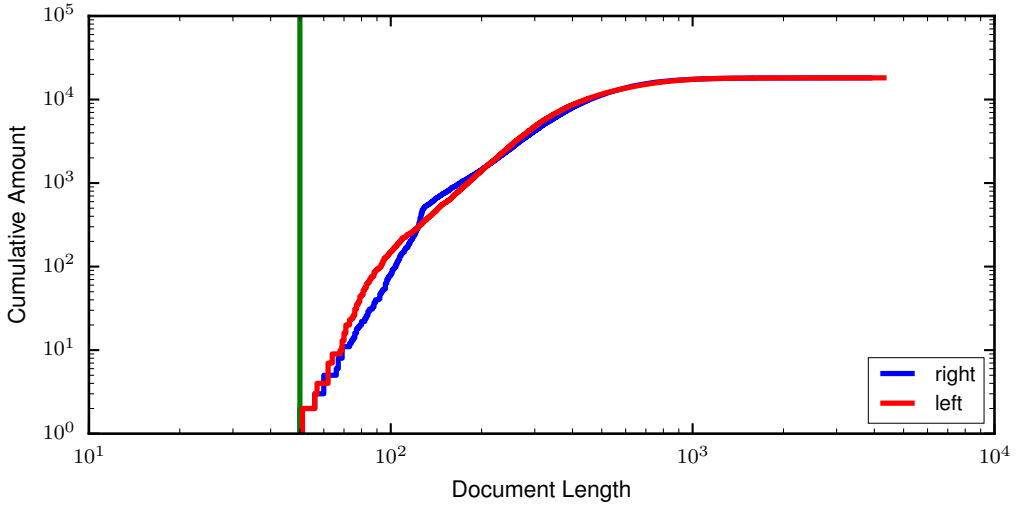


Figure 14: Article length distribution of left and right UK articles. The red graph shows the left oriented articles while the blue graph shows the right oriented articles. The green line represents the minimum length of 50 words imposed on all articles. Right articles show a sudden rise in documents of length 125 to 150 words, probably caused by length threshold of a news source.

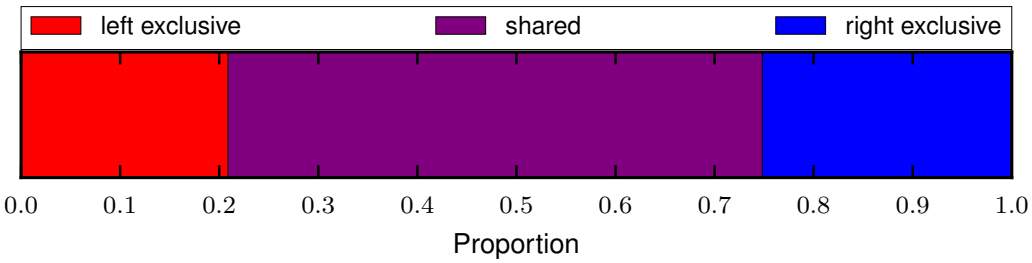


Figure 15: Unique and Shared words in the UK corpus. The red bar shows words that appear only in the left group, the blue bar shows the words that only appear in the right group. The purple bar shows the words that appear in both groups. Compared to Wikipedia data set, significantly bigger proportion of shared words. Right documents show a slightly larger exclusive vocabulary than the left documents.

Figure 15 shows the vocabulary distribution of the stemmed UK corpus. In contrast to the Wikipedia data set the Event Registry articles share the same language. Therefore, the articles share a lot more words and the difference in word diversity is not as large as in the Wikipedia data set. There were 79,813 unique words used in the stemmed UK data set. The left articles used 59,715 of these words while the right articles used 63,145 unique words. 43,047 words appear in both vocabularies. On average a pair shared around 117 words which almost triples the amount shared in an average Wikipedia pair.

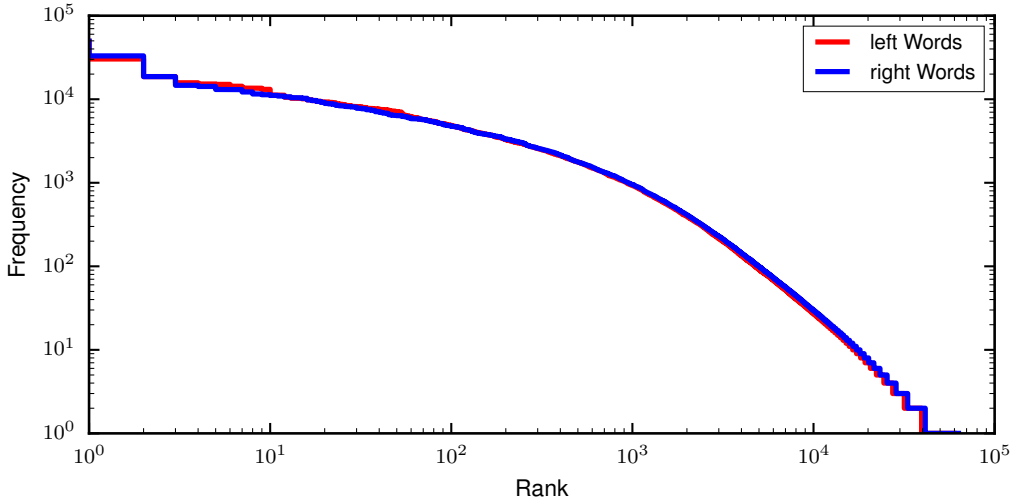


Figure 16: Word frequency of left and right articles from the UK. The red graph shows the left words, the blue graph represents the right words. The word frequency per rank is very similar as both groups use the same language.

Figure 16 shows the word frequency relative to the words frequency rank. In contrast to Figure 12 the word frequency distribution is almost identical for both groups as they share the same language. The word frequency per rank of both groups are very close together and almost overlap. There are no blatant inconsistencies between both distributions.

4.4.3 Event Registry US Data Set

The US corpus is the smallest of all three corpora. Even though there is a higher amount of different sources during the extraction, the corpus contains the least amount of pairs. The US corpus contains 4,721 pairs and consists of articles with left and right political orientation. A left article contains on average 514 words while a right article has on average a length of 468 words.

The source distribution across all pairs have been depicted in Figure 17. The most frequent left source was the "Washington Post". The most frequent right source was

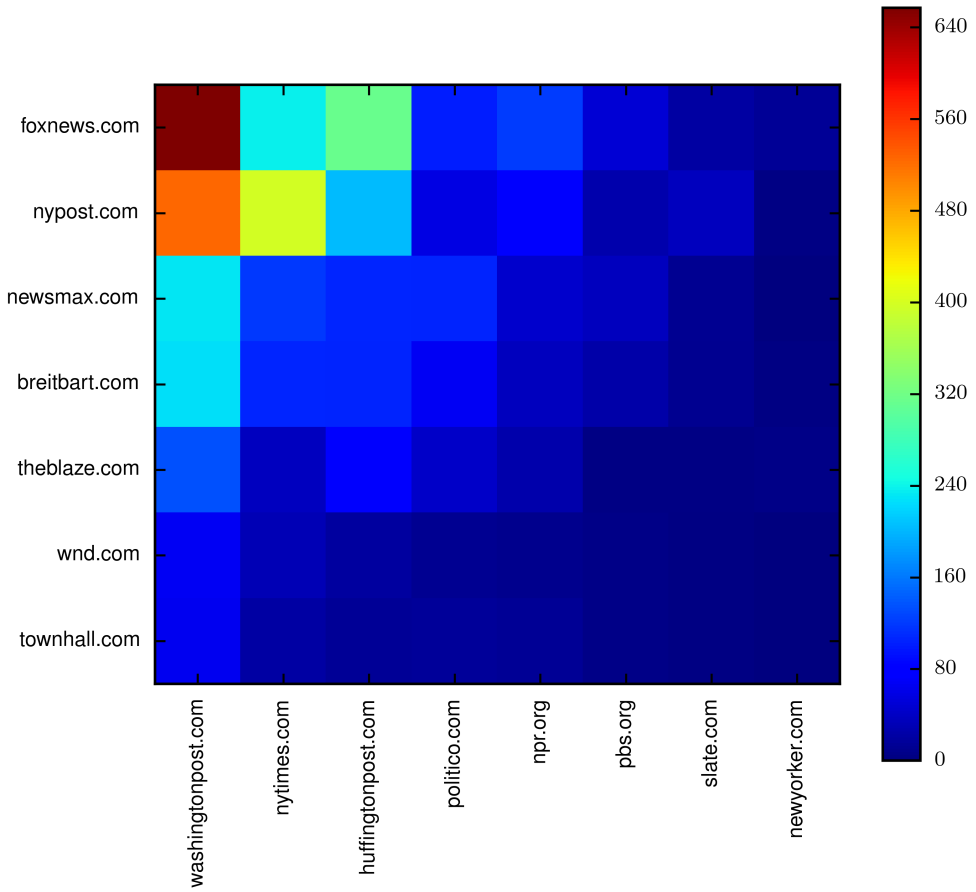


Figure 17: Source distribution of left and right US articles. Red shows a high frequency of a pair. Blue indicates a low frequency. "Washington Post" and "Fox News" are the most common pair.

"Fox News" closely followed by the "New York Post". It is interesting that "New York Post" gets paired a lot more frequently with the "New York Times" than the most frequent source "Fox News". This can probably be explained with the locality of both newspapers.

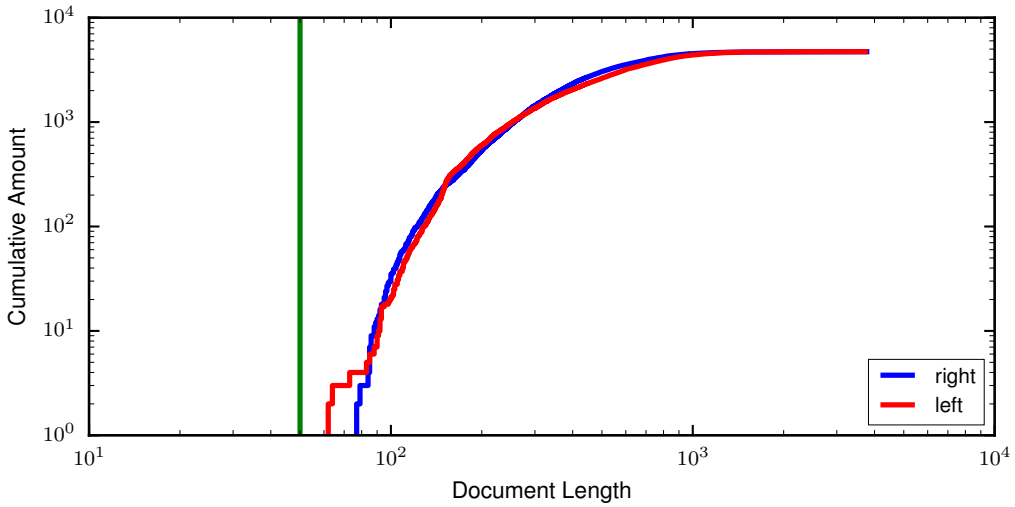


Figure 18: Article length distribution of left and right articles from the US. The red graph describes the left articles. The blue graph describes the right articles. The green line shows the minimum article length of 50 words. This time, left articles show a sudden rise in documents of length 125 to 150 words, probably caused by length threshold of a news source.

The cumulative article length distribution is displayed in Figure 18. Similar to the UK corpus, the distributions are similar and do not show a lot of difference across groups. This time the left documents show a quick rise in documents at the 125 to 150 word mark. Again, this might be explained by a policy of one or multiple news sources and was not investigated further.

When comparing the articles in a pair, a left article is longer in 58.5% of all cases while a right article is longer in only 41.2%. Compared to the UK data, where the right articles were longer, the situation has turned and is more pronounced. In the average pair, the left article is longer by about 10%.

Figure 15 shows the vocabulary distribution of the stemmed US corpus. Overall there are 41,321 unique words in the articles of the US data set. This amount cannot be compared to the UK data, because of the smaller sample size. The left articles used 33.177 different words while the right articles used 30,655 different words. Both groups share 22,511 of their vocabulary. This time the left articles show a larger vocabulary over the right articles. Similar to the UK pairs, the articles of a pair have on average 112 words in common.

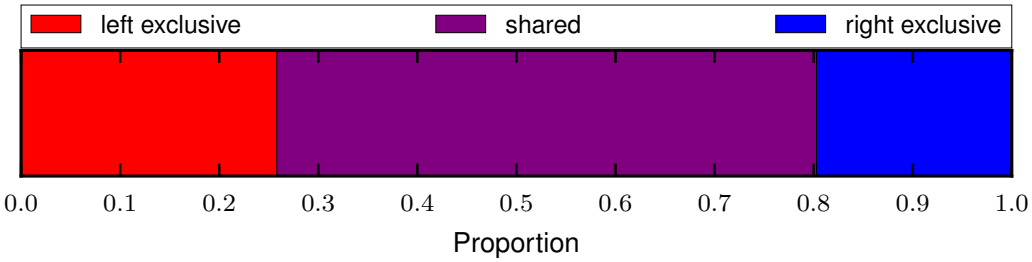


Figure 19: Unique and Shared words in the US corpus. The red bar shows words that appear only in the left group, the blue bar shows the words that only appear in the right group. The purple bar shows the words that appear in both groups. Similar to the UK corpus, most words are shared across groups. The amount of exclusive words in the left group is higher than the amount of exclusive words in the right group.

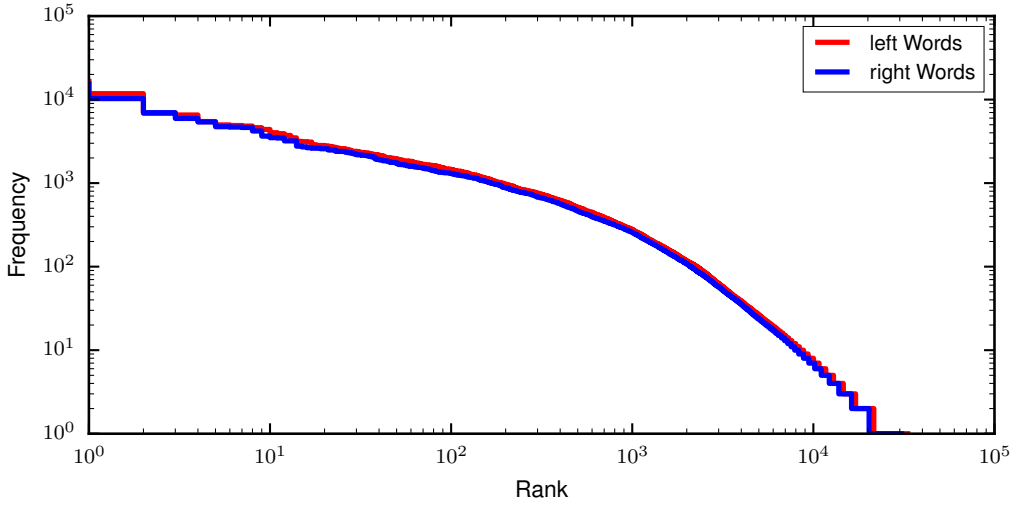


Figure 20: Word frequency of the left and right US articles. The red graph shows the left words, the blue graph represents the right words. Equal to the results of the UK corpus, the word frequency per rank is very similar as both groups use the same language.

The word frequency distribution is displayed in Figure 20. The observations match the observations of the UK word frequency plot. As both groups use the same language the word frequency distribution is very similar and almost overlaps. There are no obvious inconsistencies.

5 Experimental Results

This section presents the results of the four different experimental settings. The settings were explained in section 3.5. They give answers to the four lead questions posed in section 1.1.

During the experiments multiple LDA models were built with three different topic parameter settings – 64, 128 and 256 topics – to ensure reproducibility of results. The resulting models represent a small scale, medium scale and large scale model trained on these training corpora. Thus, each experiment produces three different plots. The perplexity range varies depending on amount of topics LDA learns due to gensim's perplexity implementation. They are not directly comparable. Hence, each model is evaluated in its personal scope of values.

Before examining the actual results, this section will give insights into the meaningfulness of the test results. To evaluate if the trained topic models predict test documents based on a sufficient amount of words, the vocabulary mismatch is calculated. It shows if a test corpus contains a high share of words that are unknown to the topic model. The prediction of words in the test phase is based on words the model saw during the training phase. Unseen words would have a very low probability, but are excluded in the experiments. Therefore, if a model only knows 10% of the words of test documents but it can assign them well, the prediction quality in the experiment is good but the result is rather flawed.

5.1 Vocabulary Mismatch

To ensure that the performance measure perplexity – which only uses known words from the training corpus – is meaningful, the fraction of words in the test corpus that have not been captured in the training corpus are examined for each group. This fraction will be called *vocabulary mismatch*. A high vocabulary mismatch indicates that a large proportion of words used in the test documents is unknown to the model. The topics have been assigned based on a few known words, while discarding most of the actual content. This is mainly important for the Wikipedia Corpus as it contains two different languages.

Figure 21 shows the vocabulary mismatch relative to the proportion of English documents in the training corpus. English documents have on average a lower vocabulary mismatch in their documents than German documents. Even if only 10% English documents are contained in the training corpus the test corpus only has a mismatch of 25%. At this point the German proportion of documents is at its highest point with 90%. Nevertheless, the German test documents already show higher

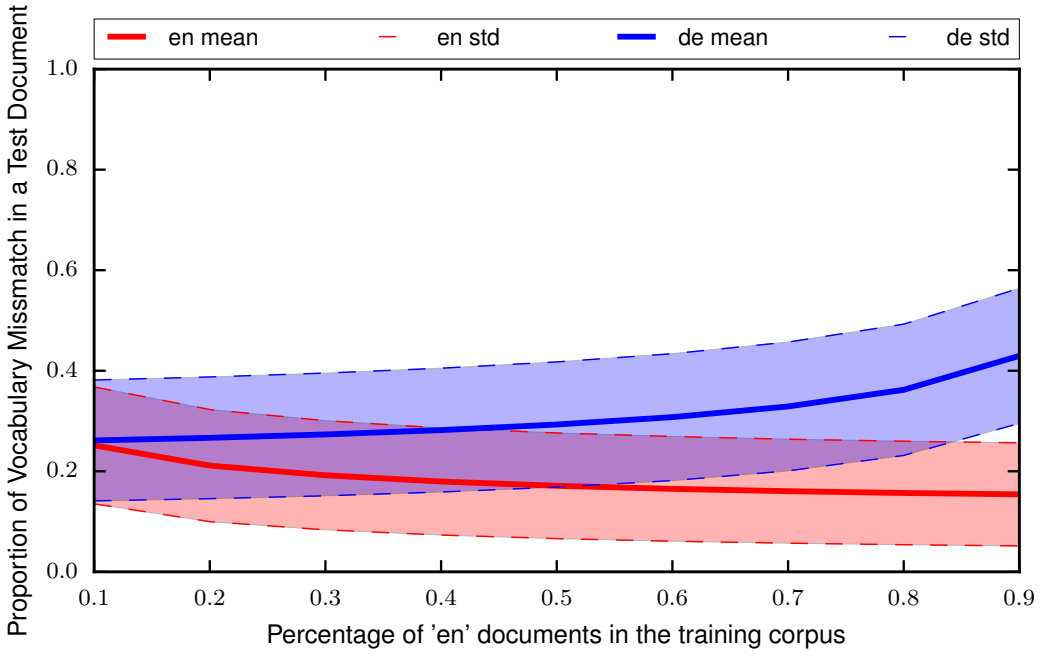


Figure 21: Average vocabulary mismatch in Wikipedia samples. The red area refers to the documents of the English test group and the blue are refers to the documents of the German test group. The continuous, middle line describes the mean proportion of vocabulary mismatch. The area between the two dashed line describes the area within the sample standard deviation of the mean. The vocabulary mismatch of a group decreases when adding documents of this group to the training corpus. With only 10% German documents in the training corpus the vocabulary mismatch almost reaches 50% and should not be increased further.

mean vocabulary mismatch. The mean vocabulary mismatch is at 26%.

When increasing the fraction of English documents in the training corpus the vocabulary mismatch of the English test documents decreases slowly. With 90% English test documents the average vocabulary mismatch of the English test documents decreases to 15%. The average vocabulary mismatch of the German documents increases to 43,0% and is at its maximum.

The German test documents contain many words that are unknown to the model. One can assume that German is a more complex language than English as it contains several different grammatical cases and compound words [32]. Figure 11 in section 4.4.1 showed that the German vocabulary is significantly larger than the English vocabulary while describing the same overall concepts. Therefore, the chance of not seeing a word in the training corpus is higher.

The amount of known words in a document is still sufficient to evaluate the performance of the topic model though, because at least half of these words are known.

Nevertheless, it is not recommended to increase the size gap between both groups any further as the mismatch will only rise and weaken the prediction results. It is worth to note that a real corpus often contains more than two languages or a very complex language such that the vocabulary mismatch will probably rise significantly in such a corpora.

The plots of vocabulary mismatch for the Event Registry corpora have been omitted. The vocabulary mismatch of both groups shows a stable mean below 10% as all documents share the same language. The mismatch is significantly lower than the vocabulary mismatch of the Wikipedia samples, such that they are well suited for inference and likelihood calculations.

5.2 Overall Perplexity

This section answers the research question **(i) if the presence of groups in the training corpus of a topic model influences its prediction performance**. It is unknown if a topic model's prediction performance on a test corpus is influenced by latent groups. There are two realistic scenarios for the effect of groups on the perplexity in the test corpus:

- The perplexity could rise: The model could mix topics of both groups together, learning topics which never occur within a single document. It also could learn topics for the majority group, and neglect the minority which is most important for a good model fit. In this case it would be necessary to split the corpus based on these groups to achieve better results.
- The perplexity could not change at all, if the topic model would be unaffected by groups – group-specific topics could be detected proportionally to the relative share of the group. This would indicate that a topic model can be trained on imbalanced corpora without special treatment of the groups, without affecting the topic quality.

This section evaluates the perplexity over the complete test corpus relative to each relative group size. The complete test corpus contains held-out words from all test documents that belong to the training documents used to train the topic model at the given relative group size.

5.2.1 Wikipedia

Figure 22 describes the median perplexity development over all test documents of the Wikipedia samples relative to the fraction of English documents in the training corpus. For all tested models, the perplexity is higher when trained on a lot of German documents and decreases monotonically when more English documents are added. At 90% English documents the model can predict the test documents best.

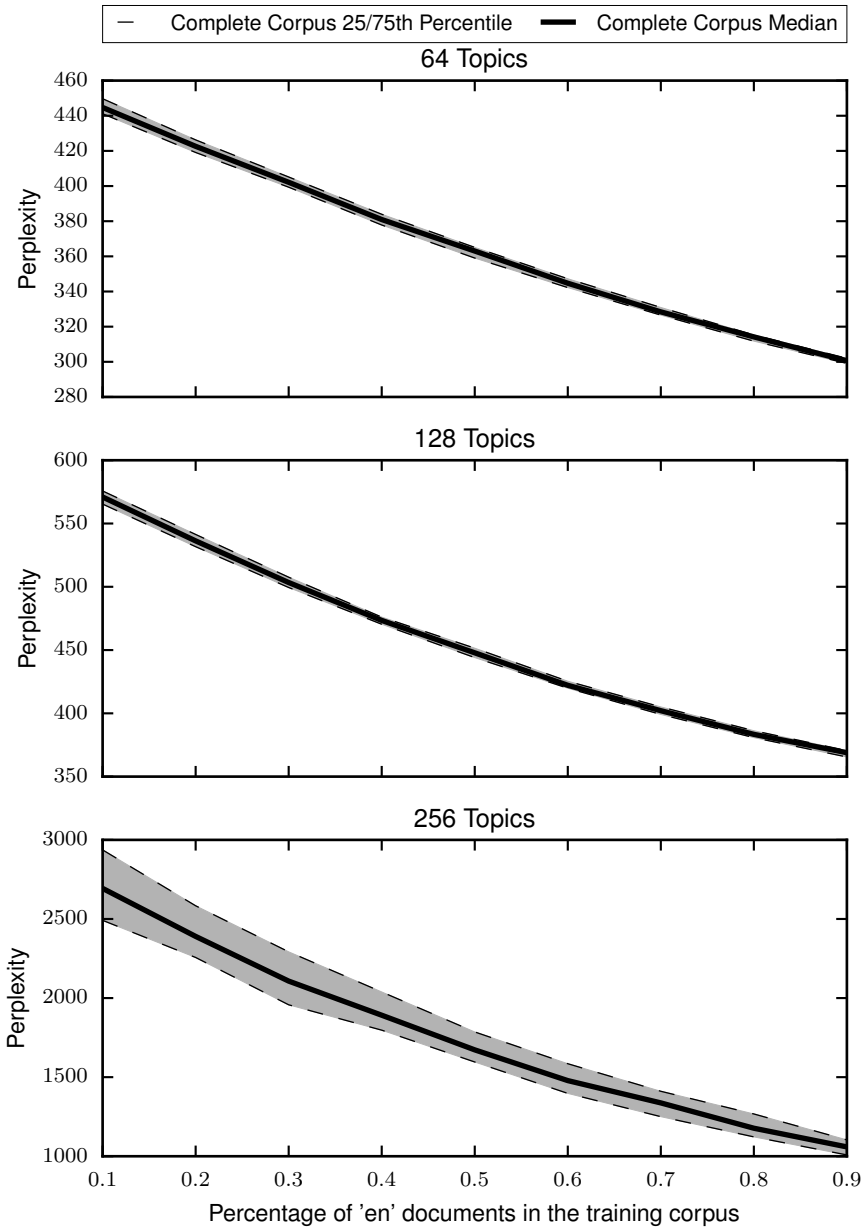


Figure 22: Perplexity over all test documents of the Wikipedia samples relative to the relative English group size. The continuous black lines describes the median perplexity of the test corpus. The dashed black lines describe the 25th and 75th percentiles. Perplexity is at its peak with 90% German documents and monotonically decreases when adding English documents. The perplexity adapts to the perplexity of the 100% corpus of each group.

This underlines the assumption made in section 5.1 and section 4.4.1 that German is a more complex language than English. Its vocabulary contains more unique words overall. Consequently, the model struggles more when predicting the German test corpus.

The experiment on the Wikipedia corpus shows that the **presence of latent groups in training corpora does influence the prediction performance** of a standard topic model. The overall prediction performance **depends on how well each group can be predicted and how big its share is**. A corpus with a majority of German documents is harder to predict than a corpus with a majority of English documents. The mixture proportion of both groups controls the overall perplexity.

5.2.2 Event Registry UK

Figure 23 describes the perplexity development over all test documents of the UK corpus relative to the fraction of left documents in the training corpus for LDA models. All graphs of LDA trained on different topic sizes show different results. The mean perplexity of LDA with 64 and 128 topics barely changes when adding or removing left documents. The mean perplexity of LDA with 256 topics changes significantly more. But when taking into account the overall scope and inter-percentile range of perplexity values of this model, the values stay rather stable as well.

The results of the experiment on the UK corpus **do not show an influence of groups on the overall prediction performance**. All graphs move rather **unpredictably**. The only common feature is that a 90% right corpus is slightly worse to predict than a 90% left corpus. But the effect is rather insignificant as the perplexity between those extremes varies too much. Hence, it is safe to assume that the prediction performance of these samples **depends on other influences** than the presence of groups.

5.2.3 Event Registry US

Figure 24 describes the perplexity development over all test documents of the UK corpus relative to the fraction of left documents in the training corpus. LDA models that learned 64 or 128 show a declining trend when increasing the amount of left documents. This trend cannot be found in LDA when learning 256 topics. The mean perplexity in LDA with 256 topics stays rather stable at a very high perplexity value but can deviate a lot. This effect might be caused by the topic model learning too many topics on a rather small sample (4000 docs), such that it cannot reasonably assign the words to topics. Hence, the results of LDA with 256 topics are disregarded for now.

The results of the experiment on the US corpus show a **minor influence of groups on the overall prediction performance**. Figure 19 in section 4.4.3 shows a small overhead in vocabulary size for the right articles. So, similar to the declining trend in the Wikipedia samples, the **declining trend** in overall perplexity can be explained with the **vocabulary size of the group**. The right articles use more unique words in

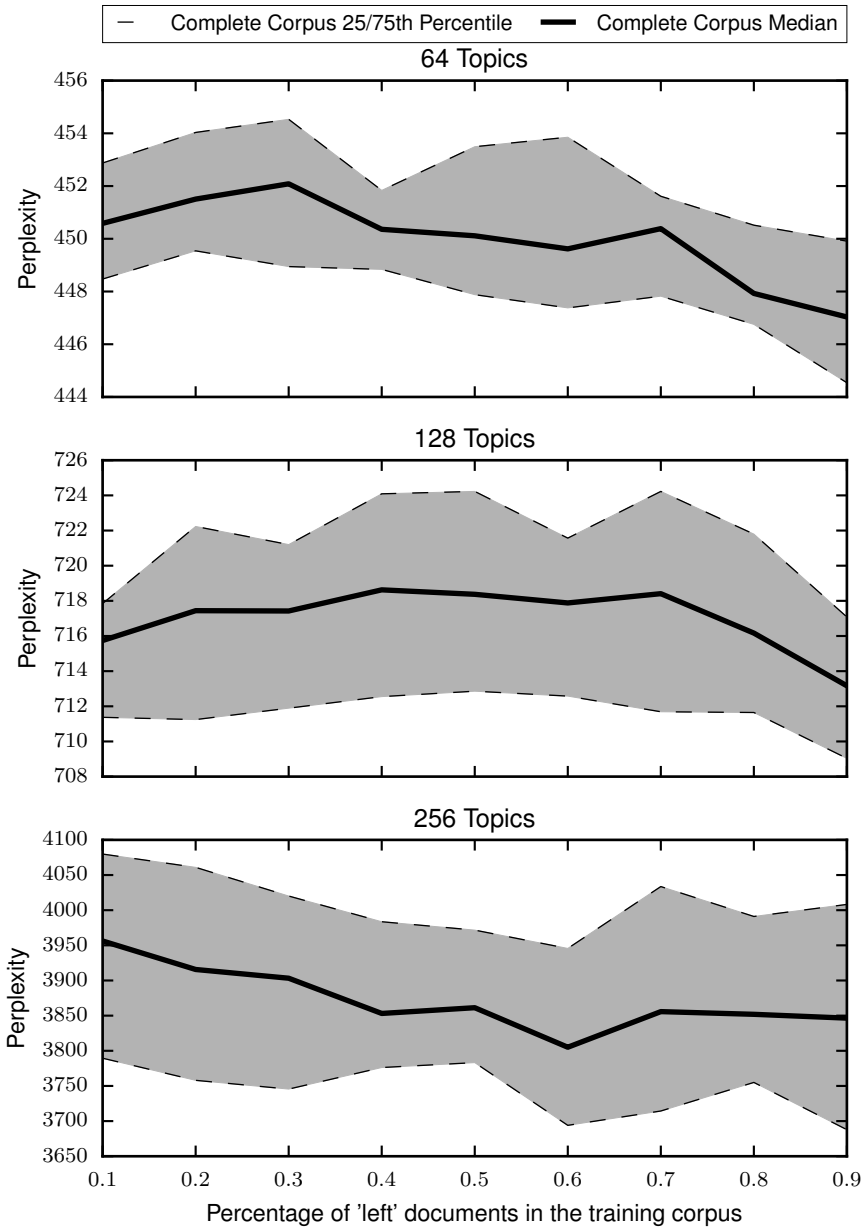


Figure 23: Perplexity over all test documents of the UK samples relative to the left group size proportion. The continuous black lines describes the median perplexity of the test corpus. The dashed black lines describe the 25th and 75th percentiles. Perplexity does only vary in a small perplexity window – besides 256 topics, the perplexity varies more and the inter-percentile distance is larger. There are no common patterns between all three topic sizes.

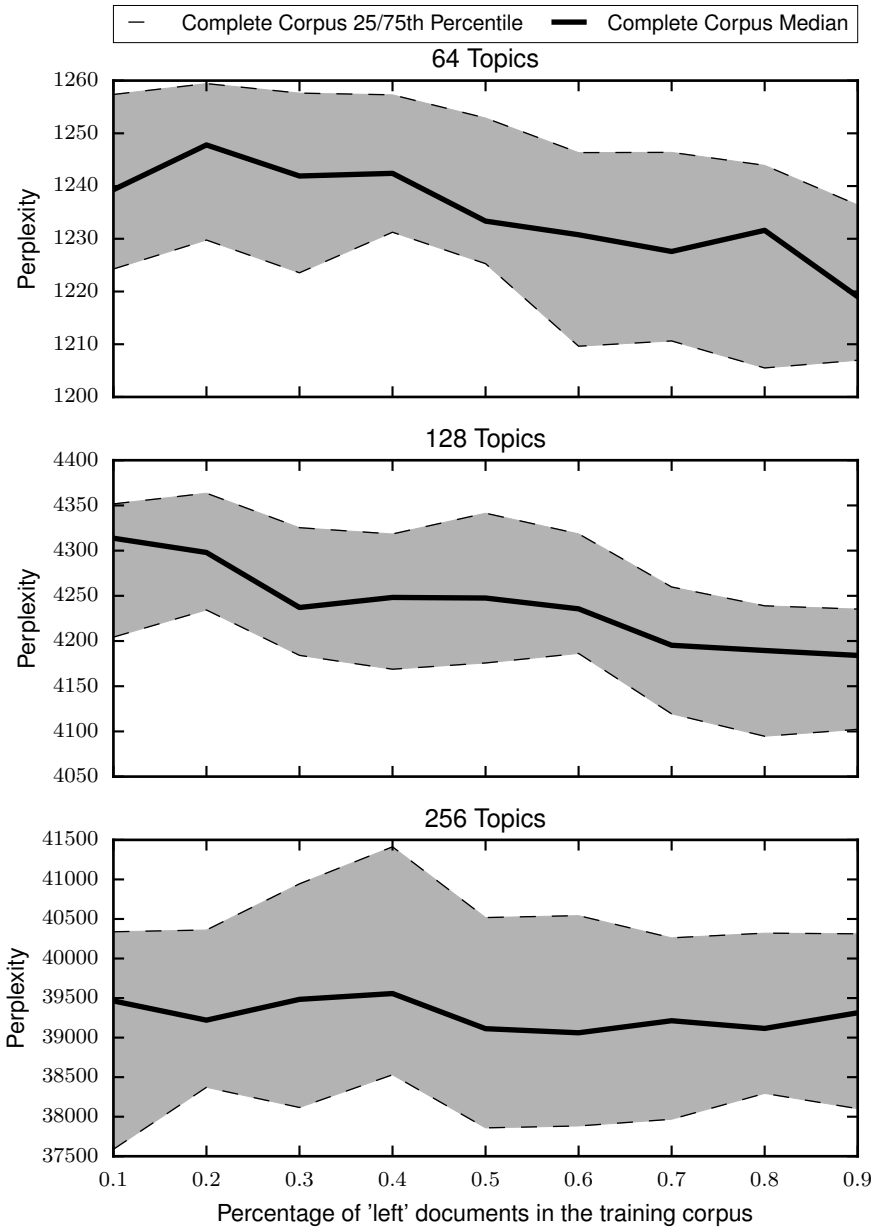


Figure 24: Perplexity over all test documents of the US samples relative to the left group size proportion. The continuous black lines describes the median perplexity of the test corpus. The dashed black lines describe the 25th and 75th percentiles. Trend of decreasing perplexity when increasing the share of left documents when training on 64 and 128 topics. Not as pronounced as in the Wikipedia samples. Results of 256 topics rather stable but with a larger inter-percentile range.

their documents, therefore they are harder to predict for the topic model and the overall perplexity is higher.

5.3 Group-Specific Perplexity

This section will answer **(ii) if the group-wise prediction performance changes when the relative group size relation in the training corpus of a topic model is varied.**

It is reasonable to expect that the prediction performance of a group suffers when the group is pushed further into a minority position, because the minority group's share in the training corpus is decreasing. There might even be a point of imbalance at which a topic model can not cover the minority group properly anymore, because the group's share of documents in the training corpus is too small. It is unknown when and if this point occurs.

To test this question, this section evaluates the perplexity over each group's test corpus relative to the relative group size.

5.3.1 Wikipedia

Figure 25 shows the perplexity development over each groups' test documents in the Wikipedia corpus relative to the fraction of English documents in the training corpus. In almost all cases, English documents can be predicted better than German documents. This gap can be reasoned with each languages complexity. Even though English documents are on average longer than German documents, the German documents use more unique words (see section 4.4.1). Therefore, a sufficient amount of German documents is necessary to build topics that can predict the test documents well. English corpora can build topics with better prediction performance with a lower share of documents because English documents use less unique words.

When increasing the amount of English documents in the training corpus, the perplexity on English test documents decreases monotonically. At the same time the perplexity of German documents increases monotonically. The English documents show an increase in perplexity when only 20% or less English documents are in the training corpus, the German documents show a similar increase when only 40% or less are German documents in the training corpus. The increase in perplexity can be explained by the fading amount of the minority language's words in the sample vocabulary.

The results when training on 256 topics differ from the other results. It is the only result within the test range where the English documents can be predicted worse than the German documents at one point. A possible explanation of why this case occurs within the test range only when training 256 topics is that the other models could not keep track of enough topics to increase the prediction performance of German documents to the level of English documents. When more topics can be trained, the more complex German documents can be covered more appropriately.

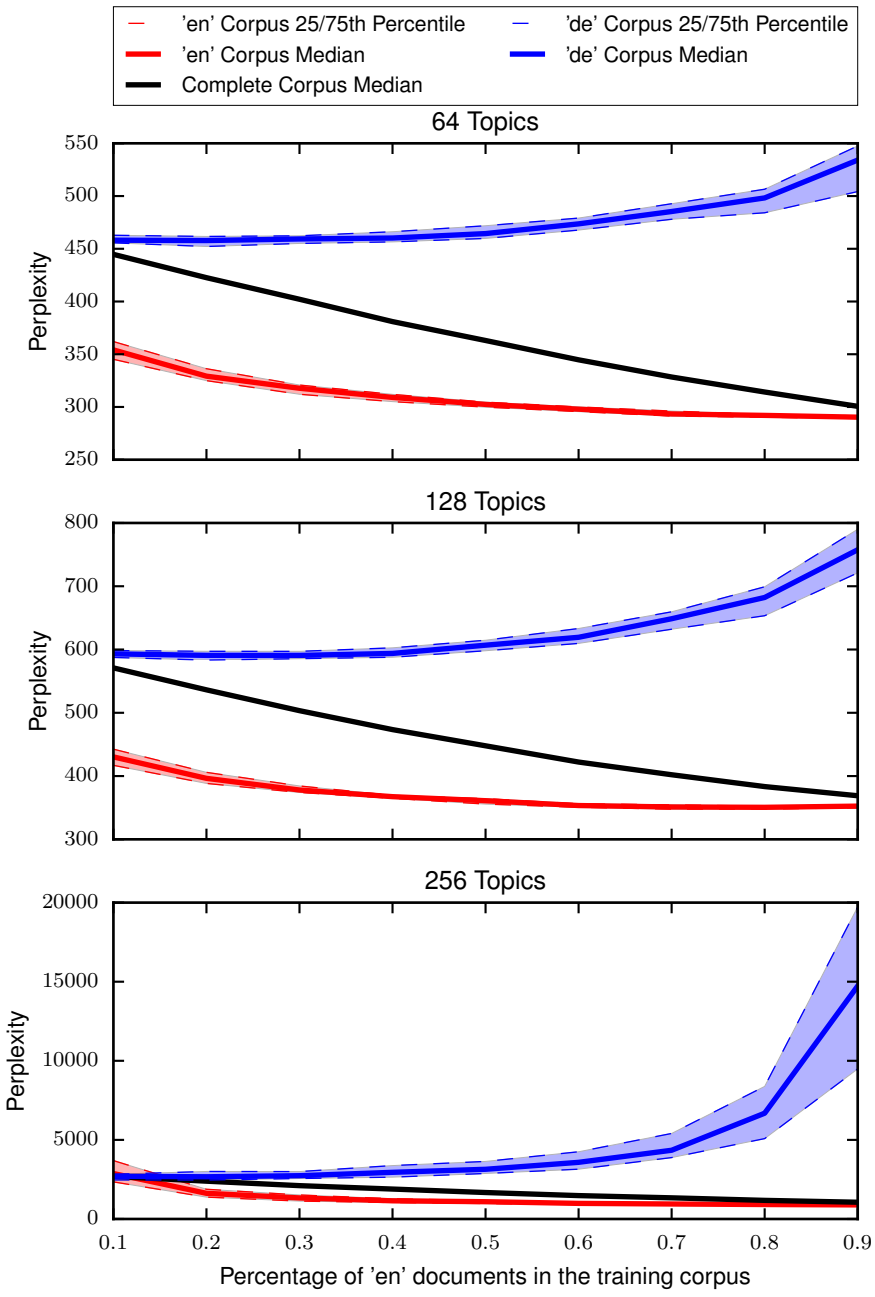


Figure 25: Perplexity over test documents of the Wikipedia samples per group. The red line shows the median perplexity of the English test corpus, the blue line shows the median perplexity of the German test corpus. The black line shows the perplexity of the complete test corpus (from Figure 22). The dashed line shows the 25th/75th percentile of perplexity of each test group. In general, German documents are harder to predict than English documents. English documents show a significant perplexity increase when they provide 20% or less of the training corpus. German document show a perplexity increase when providing less than 40% of the training corpus. The rate of growth increases with decreasing share of documents.

This case occurs when only 10% of the test corpus consist of English documents. It is possible this case occurs in a range below 10% for models trained on less topics – as long as the vocabulary mismatch does not prevent such a prediction. These cases have not been investigated during this thesis.

In general, all results show that the prediction performance is not solely dependant on a minority position of a group. It is true that the prediction performance worsens when pushed further into a minority position, but the graphs demonstrate that the English minority group frequently shows a better prediction performance than the German majority. The amount of topics learned influences the relative prediction performance between both groups.

There have to be additional factors that influence the predictability besides the relative group size. As already assumed in section 5.2 the complexity of a language is a reasonable explanation. German is a more complex language [32], such that the topic model cannot predict it as well as the English documents even if German documents are the majority during training.

5.3.2 Event Registry UK

Figure 26 shows the perplexity development over each groups' test documents in the UK corpus relative to the fraction of left documents in the training corpus. The perplexity on right test documents shows a decreasing trend when increasing the amount of left documents in the training corpus when LDA trained 64 or 128 topics. At the same time the perplexity of right documents show an increasing trend. The perplexity decreases faster for left documents than the perplexity of right documents rises.

When LDA is trained on 256 topics, both perplexity values are almost equal. The graphs are barely interpretable. It shows an irregularity which is not occurring in LDA trained with 64 or 128 topics. It is reasonable to assume that the increased amount of topics is sufficient to train topics that cover both groups, such that the perplexity equalises.

When LDA is trained on 64 topics both groups almost have the same perplexity value when the training corpus contains 60% left documents. When trained on 128 topics this point moves towards 90% left documents in the training corpus. Before this point, right documents show a better prediction performance, beyond this point left documents show a better prediction performance. Similar to the Wikipedia corpus, the break-even point of perplexity is influenced by the amount of topics.

The two graphs of LDA with 64 and 128 topics show that, in a same language corpus, the prediction performance of each group suffers if forced into a minority role. When there are are enough topics to learn e.g. 256 topics, this effect slowly nullifies. It is worth to note that the test sets of both groups still show different perplexity values when trained on lower amounts of topics. This behaviour is surprising, as it exhibits language differences beyond language or covered topics (which were controlled for).

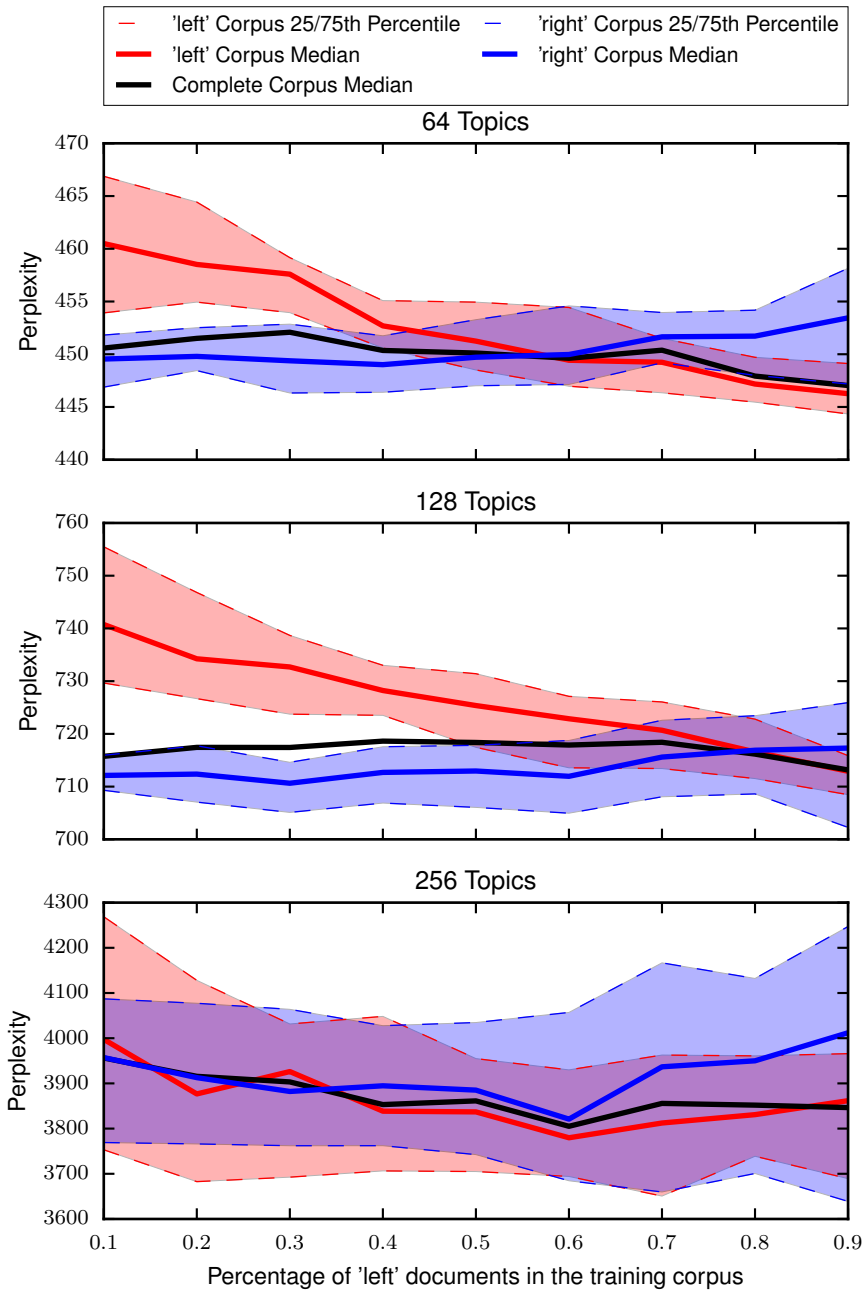


Figure 26: Perplexity over test documents of the UK samples per group. The red line shows the median perplexity of the left test corpus and a blue line shows the median perplexity of the right test corpus. The dashed line shows the 25th/75th percentile of perplexity of each groups test corpus. The black line shows the perplexity of the complete test corpus (from Figure 23). Models trained on 64 and 128 topics show similar results. Perplexity on left documents decreases when adding left documents to the training corpus. Perplexity on right documents increases, but at a slower rate. Models trained on 256 topics show a similar perplexity for all documents without showing common patterns with the other two models.

5.3.3 Event Registry US

Figure 27 shows the perplexity development of each groups' test documents in the US corpus relative to the fraction of left documents in the training corpus. The results differ from the results of the UK and Wikipedia corpus.

Both groups' perplexity increases monotonically with increasing amount of left documents in the training corpus. The rate of growth is almost equal for both groups. It is worth to note that this keeps the perplexity of the complete corpus at a stable level.

Right articles always show a larger perplexity than left articles. The reason behind this difference is currently unexplained. Different to the Wikipedia corpus, left articles – which are easier to predict – show a larger vocabulary. It is possible that certain words in left articles are used together more consistently such that the left-oriented topics are better at predicting new articles. The offset of both groups shows though, that both groups have internal properties that influence the prediction performance differently.

The fact that both groups perplexity increases monotonically and at the same rate when increasing the amount of left documents in the training corpus indicates that these samples are not influenced by minority-majority relations at all. If a topic model treats both groups fairly, then there would be two lines parallel to the x-axis, because for each relative group size relation the perplexity will stay stable. These graphs are parallel, but are rising with increasing left document share.

This means that there is an additional hidden parameter that changes when adding left documents. A reasonable parameter – given the corpus statistics in section 4.4.3 – is the average document length. It shows a significant difference between 514 words in left articles and 468 words in the right articles. So, when adding left articles the average document length in the sample rises and the documents of both groups in the sample are harder to predict for LDA.

The perplexity of the complete test corpus stays stable throughout this process because the right test documents at 90% right documents are predicted almost as good as the left test documents at 90% left documents. The change relative group size relation then counter-acts the overall increase in perplexity such that it stays at an overall stable level.

5.4 Topic Assignment per Group

This section gives the answer to the question **(iii) if a topic model can distinguish between two latent groups in the training corpus**. It is unclear if a topic model trained on a imbalanced corpus builds topics that are used especially for one group or if the topics are build for both groups.

This section evaluates the proportion of topics in a topic model that can be assigned to one of the groups relative to each relative group size. A topic T is assigned to a group G if the formula $P(G|T) > 90\%$ holds.

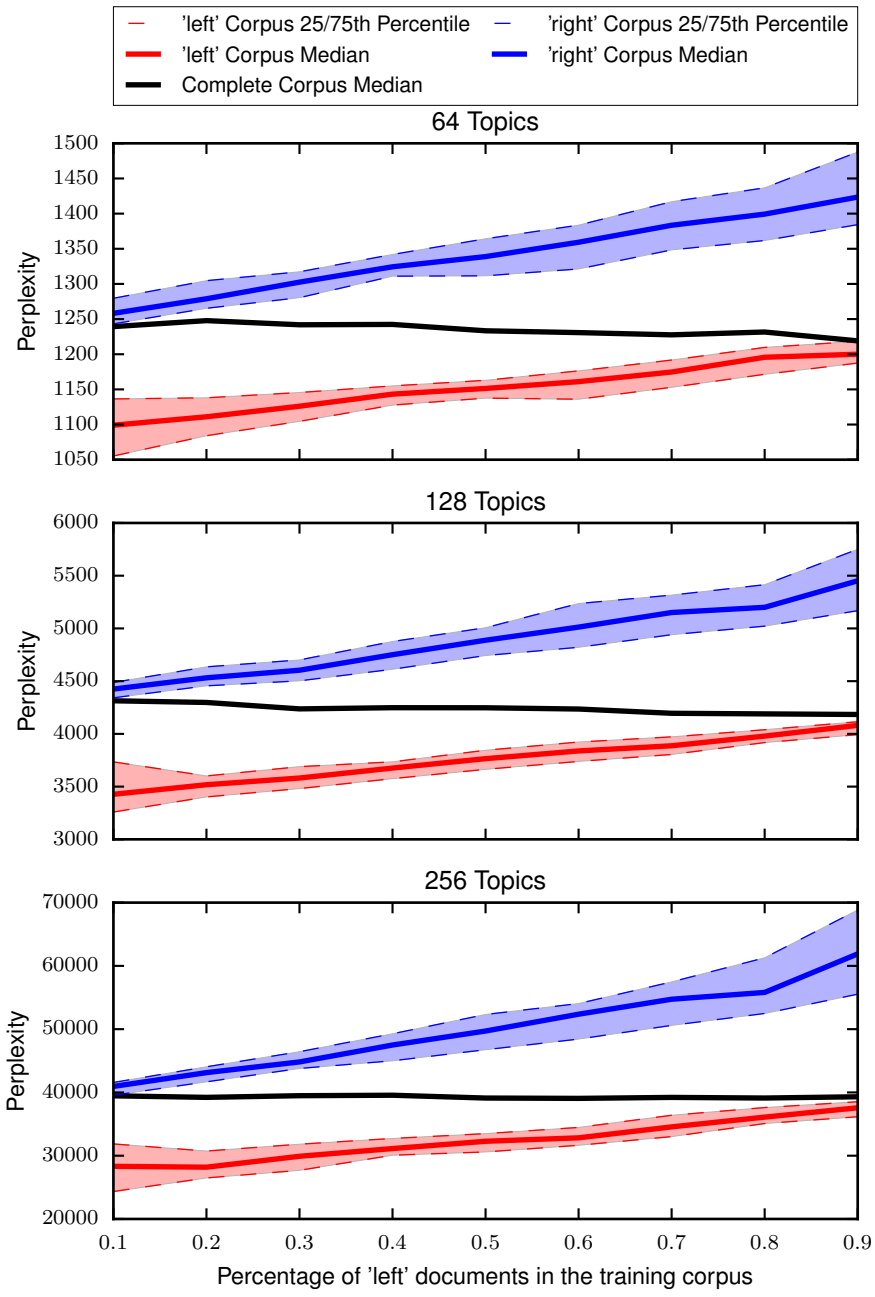


Figure 27: Perplexity over test documents of the US samples per group. The red line shows the median perplexity of the left test corpus and a blue line shows the median perplexity of the right test corpus. The black line shows the perplexity of the complete test corpus (from Figure 24) The dashed line shows the 25th/75th percentile of perplexity of each groups test corpus. Both group's perplexity monotonically increases with the share of left documents in the training corpus. Both change almost analogously, keeping the median perplexity of the whole test corpus stable.

If a topic can be assigned to a group unambiguously, this section also answers **(iv) if the proportion of group-specific topics is under-proportional to the proportion of topics in the training corpus**.

If a topic model builds separate topics for each group, it is unknown if a topic model builds e.g. 70% of its topics for a group, if this group provides 70% of the training corpus. It is expected that the amount of topics built is under-proportional, as the assignment rule is set rather strictly, but the extent of under-proportionality is unclear.

5.4.1 Wikipedia

Figure 28 shows the topic assignment for LDA trained on the Wikipedia corpus. As all graphs are similar, this section focuses on the data of LDA trained on 128 topics. **LDA trained on Wikipedia samples can distinguish both groups**.

At 10% English documents in the training corpus there are about 83% German topics and below 3% English topics. The case is reversed when there are 90% English documents in the training corpus, there are 80% English topics and 3% German topics. As the corpus gets more balanced the amount of topics that can not be assigned to one group rises quickly.

At the peak, 51% of all topics cannot be assigned unambiguously. That means in the balanced training corpus at 50% English documents only about 20% to 25% of topics can be assigned to a group. It is surprising that in the balanced corpus less than 50% of the topics can be assigned to one of the groups even though it contains two separate languages.

This refutes the claim of Boyd-Graber et al. [5] – mentioned in section 2.5.2 – which stated that LDA, when trained on poly-lingual corpora, learns topics that belong to a single language only. This thesis is the first to show, that the learned topics are used by both corpora. Even though the most frequent terms appear to share the same language, these **topics cannot be assigned to one language unambiguously as both corpora are using these topics**.

Nevertheless, the usage of shared topics is usually not desired in a topic model. Thus it is advised to not use LDA on poly-lingual corpora and to either split the corpus by language to achieve clean topics or to use a poly-lingual model [5, 18]

To further analyse the connection between the proportion of group assigned topics and the proportion of a group in the training corpus, their relation is depicted in Figure 29. It shows that minority groups are always significantly **under-represented**: The share of group-specific topics grows relative to the share of documents of a group in the corpus. As seen in Figure 29, **the perplexity of small minorities of 10% or 20% of the corpus is significantly worse than the perplexity of the majority group**. For minorities of 30% and larger, this effect is considerably smaller.

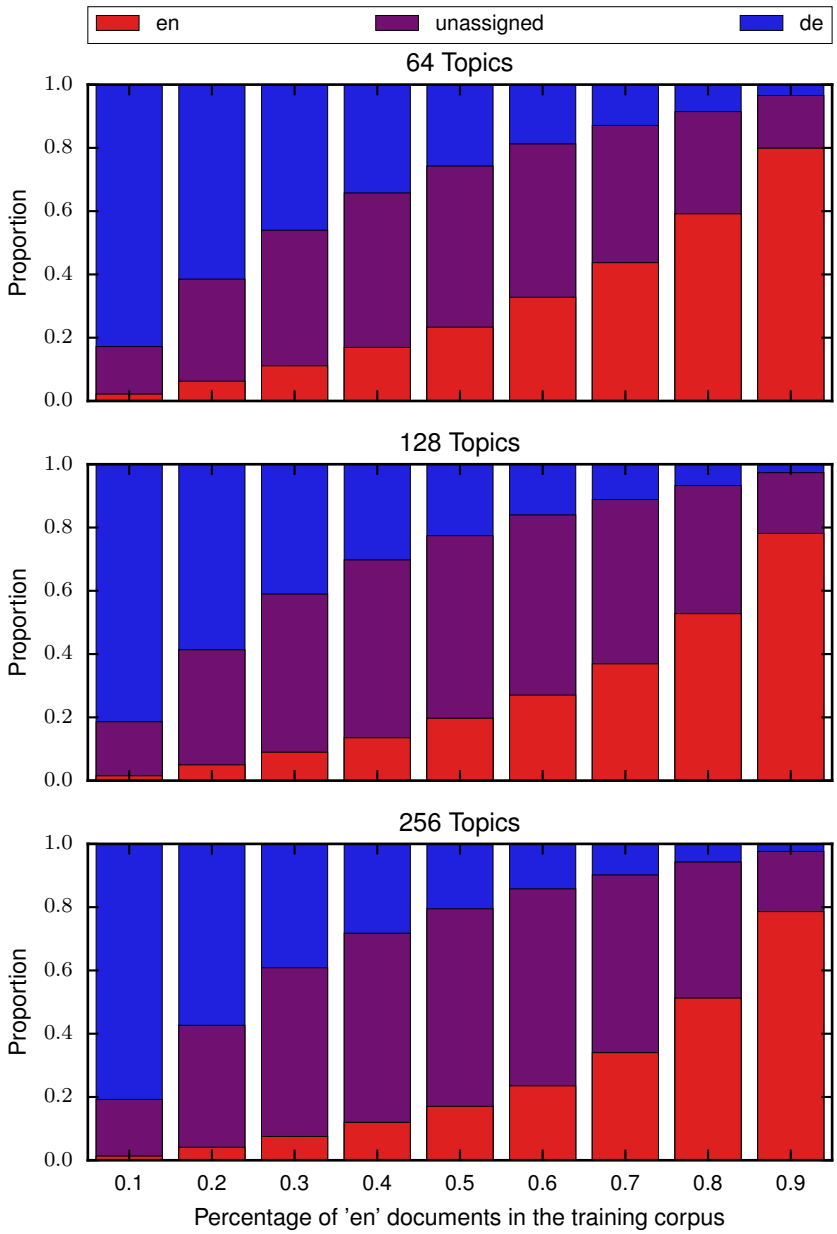


Figure 28: Topic Assignment per group in the Wikipedia corpus. Red bars show the mean proportion of topics assigned to the English group while the blue bars show the mean proportion of topics assigned to the German group. Assigned topics will be called English and German topics. Purple bars show the mean proportion of topics that cannot be assigned to either group. LDA can differentiate between both groups. The amount of assigned topics per group depends on the share of this group in the training corpus. Close to the balanced state, the model creates a lot of shared topics that cannot be assigned unambiguously.

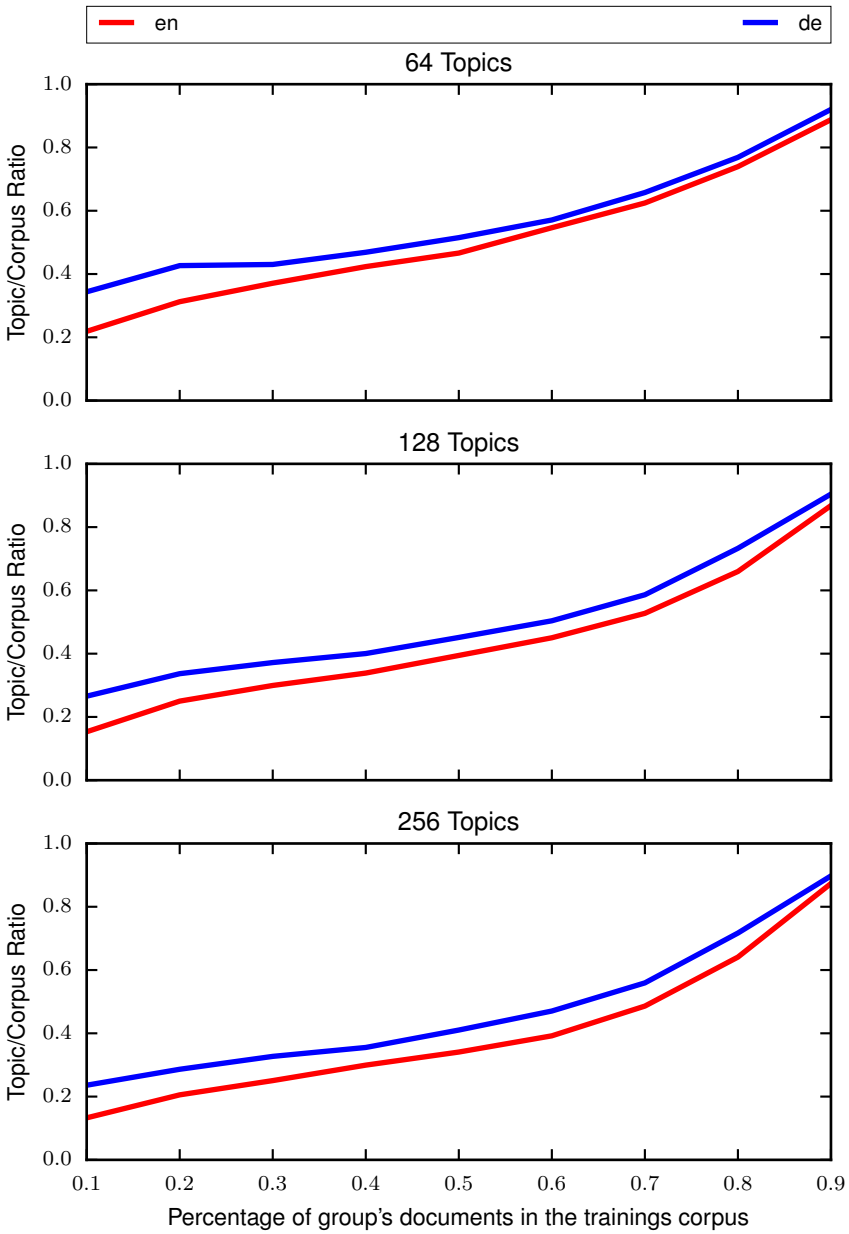
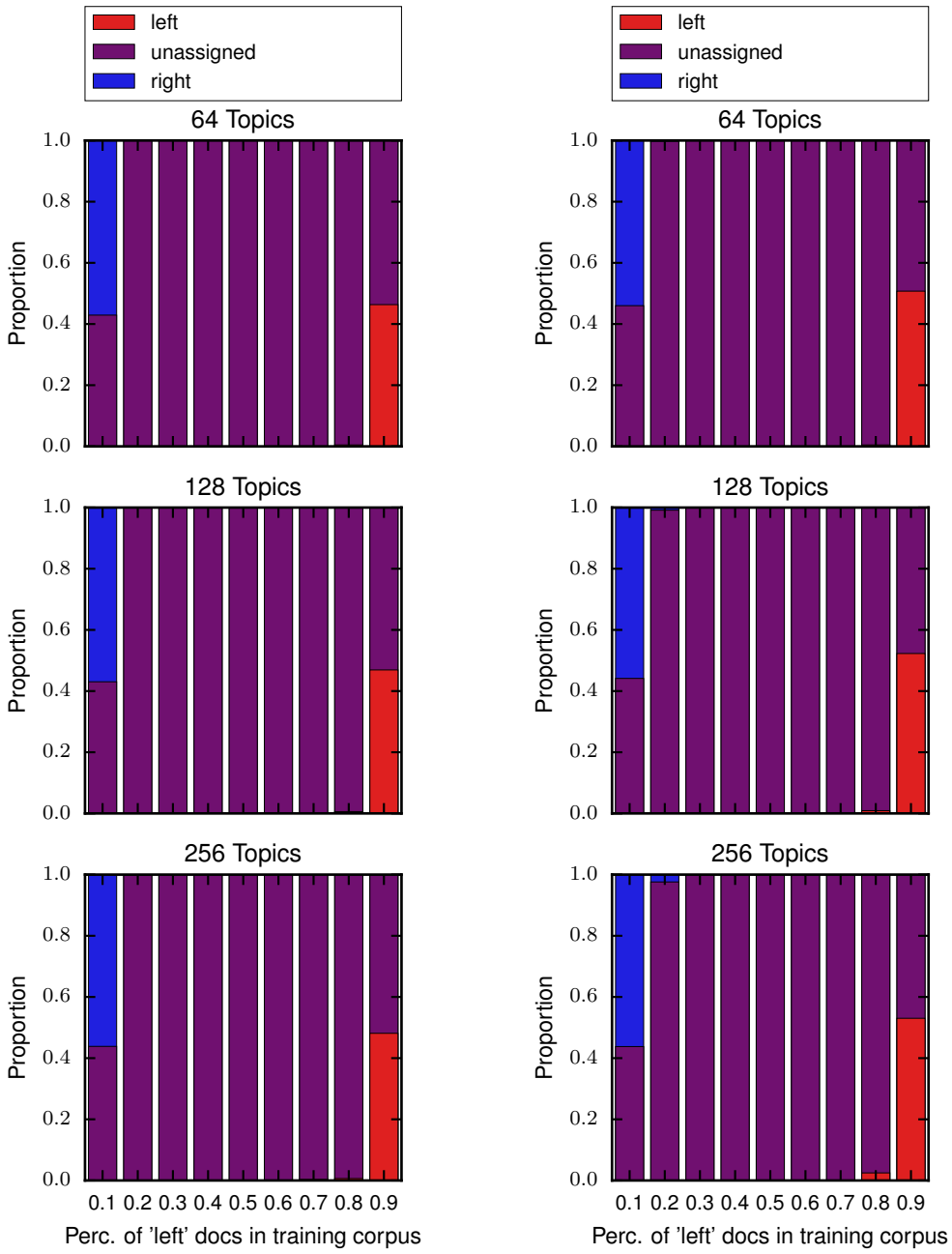


Figure 29: Ratio of the share of group-assigned topics to the share of group documents in the training corpus. Smaller minorities get represented in under-proportionally many topics. They are under-represented.



(a) UK corpus

(b) US corpus

Figure 30: Topic Assignment per group in the Event Registry corpora. Red bars show the mean proportion of topics assigned to the left group while the blue bars show the mean proportion of topics assigned to the right group. Assigned topics will be called left and right topics. Purple bars show the mean proportion of topics that cannot be assigned to either group. In both data sets almost all topics are shared. The documents that could be assigned to a group can only be assigned due choice of the assignment threshold.

5.4.2 Event Registry

Figure 30 shows the topic assignments to groups for LDA trained on the US and UK corpus. Since both corpora contain groups that do not differ by language but by political orientation, LDA generates almost no group-specific topics. In the range of 20% to 80% of English documents in the training corpus almost no groups can be assigned to a group. That means, the topic model uses these topics in both groups.

Only the corner cases of 10% and 90% English documents show a clear assignment. This assignment happens due to the interaction between the threshold and the construction of the corpus though. The threshold to assign a topic is set to 90%. Therefore, if the training corpus contains 90% documents of a group, it will automatically assign about 50% of the topics to that group by chance.

It appears that the differences between both groups are too small to clearly assign them to a group. As decent **group assignments are not possible**, an investigation of the relation between share of topic group assignment and share of group in the training corpus is impossible.

6 Conclusion

This thesis investigated the influence of an imbalanced training corpus on LDA, the most-popular topic model. The main findings of this thesis are:

(i) The presence of groups in training corpora can influence the prediction performance of topics models as measured by perplexity due to various factors, including increased group-specific perplexity scores.

(ii) The prediction performance of topic models changes for all groups when varying the relative group sizes.

(iii) Basic topic models are able to distinguish between different latent groups in document corpora to a certain extent if differences between groups are large enough, e.g. for groups with different languages.

(iv) The proportion of group-specific topics is under-proportional to the share of the group in the corpus and relatively smaller for minorities.

To achieve these findings, three different data sets were built: a Wikipedia data set, a data set with UK articles and a data set with US articles. Using these data sets, imbalanced training corpora were sampled. The imbalanced training corpora contained documents of two distinct groups. The relative group size relation was manipulated. In order to remove the influence of the semantics of sampled documents when varying the relative group size, the covered concepts in each sample were controlled for. The training corpora were used to train a LDA model. Using held-out test data of each group the topic models were evaluated regarding their prediction performance.

Both the UK and the US data sets were constructed with English speaking articles. The latent groups in this corpus differed by political orientation. The vocabulary difference across these groups is rather small. LDA (iii), (iv) **did not build topics for groups specifically** as the difference was too small.

When changing the relative size relation between these groups (i) **they did not show any clear results regarding the prediction performance** of the topic model in general. It stayed rather stable.

When investigating the perplexity of the group specific test corpora (ii) both corpora showed completely different results. In the **UK corpus, the prediction performance of a group increased when adding documents of that group**. The point where both groups could be predicted equally well depended on the amount of topics LDA learns.

In the **US corpus, the prediction performance of both groups improved when increasing the amount of right documents**. It suggests that both groups were treated fairly, but the **perplexity depends on other group specific parameter** i.e. the article length of each group.

The Wikipedia data set contains German documents and English documents. It shows a clear vocabulary difference between both groups. Therefore, the results differed from the Event Registry corpora. The Wikipedia data showed **(i) that the prediction performance varies between languages**. English documents were easier to predict than German documents – due to a significantly smaller vocabulary – such that an increasing the amount of English documents in the corpus benefited the prediction performance.

When analysing the group-specific test sets (ii) English documents showed increasing perplexity values when providing less than 20% of the training corpus, German documents showed an even higher increase when providing less than 40% of the training corpus. In General, the **English test documents showed a better prediction performance than the German test documents**. Only when learning 256 topics, 10% English documents could be predicted worse than 90% German documents. Similar to the UK results, the results of the Wikipedia data set were influenced by the chosen amount of topics LDA learns.

The language separation in the Wikipedia data set was clear enough such that LDA **(iii) could differentiate between the two groups and build group assigned topics**. The amount of topics assigned to each group increases with the proportion of a group in the training data. It did not assign all topics though. There was always a share of topics that was assigned to both groups. This share grew, if the corpus got more balanced.

The **share of topics (iv) was always under-proportional** relative to the share of documents in the training data. The effect of **under-proportionality increases dramatically for small minorities** of 10% or 20% of the corpus.

Overall the influence of minority and majority depends on the properties of the contained groups. If the group difference is rather small e.g. political orientation, then the groups do not alter the overall prediction performance of the model and the

model does not learn specific topics for each group. The influence on each groups test corpus is significant and differs though, even though the overall prediction on the test corpus might look similar.

If the group difference is large e.g. languages, then the topic model prediction performance adapts to the majorities prediction performance. That is, if a training corpus contains 80% articles that are easy to predict, then the perplexity will increase compared to a balanced corpus. The topic model even learns specific topics for each language. At the same time it is surprising that a rather large share of topics stays unassigned and is used by both language groups.

These unassignable topics contradict a claim of Boyd-Graber et al. [5]. They stated that LDA, when trained on poly-lingual corpora, would learn only language-specific topics (e.g. English and German topics). In the experiments of this thesis, the behaviour of LDA on poly-lingual corpora was examined for the first time and it could be shown that not all learned topics are language-specific. A significant share of the topics in the Wikipedia data set could not be assigned unambiguously to one language.

These findings have practical implications for the application of LDA: When working with topic models, it is necessary to test for groups with significant vocabulary differences (e.g. different languages) in the corpus. These groups should be identified, and group-specific parameters should be learned – either by using a topic model which models groups (e.g. poly-lingual topic models [5, 18]) or by learning a specific topic model for each group. This will prevent topics of mixed groups such as mixed-language topics, and circumvent majority-minority situations which can have serious impact on the predictability of the minority group.

Standard corpora, with common vocabularies or small unique vocabularies do not pose such a large problem for LDA. Even though the perplexity values differ across groups, they stay close together such that the impact is not as influential as the impact of a language difference.

The variety in all results shows that there exist additional group-related features that control the prediction performance, e.g. the complexity of a language and the length of articles in each group. It is up to further research how these features impact the prediction performance and if they enhance or reduce the effect of a minority role. Suggestions for possible features that could have an influence are: the sample size, the amount of vocabulary overlap across groups, the article length difference across groups and the complexity of the language used in a group. Especially the impact of various languages on the perplexity of topic models is interesting, because a complexity rank of languages can be created.

References

[1] David M. Blei, Lawrence Carin, and David B. Dunson. Probabilistic topic models. *IEEE Signal Processing Magazine*, 27:55–65, 2010.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Arnim Bleier. Practical collapsed stochastic variational inference for the hdp. *CoRR*, abs/1312.0412, 2013.

[4] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[5] Jordan L. Boyd-Graber and David M. Blei. Multilingual topic models for unaligned text. *CoRR*, abs/1205.2657, 2012.

[6] Kevin Robert Canini and Thomas L. Griffiths. A nonparametric bayesian model of multi-level category learning. In Wolfram Burgard and Dan Roth, editors, *AAAI*. AAAI Press, 2011.

[7] Jonathan Chang, Jordan L Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Nips*, volume 31, pages 1–9, 2009.

[8] Enhong Chen, Yanggang Lin, Hui Xiong, Qiming Luo, and Haiping Ma. Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing & Management*, 47(2):202–214, 2011.

[9] Gregor Heinrich. Parameter estimation for text analysis. Technical report, arbylon.net and Fraunhofer Computer Graphics Institute, http://www.arbylon.net/publications/text-est.pdf, August 2005.

[10] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[11] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[12] Christoph Kling. *Probabilistic models for context in social media*. PhD thesis, 2016.

[13] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[14] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68. ACM, 2009.

[15] Ying Liu, Han Tong Loh, and Aixin Sun. Imbalanced text classification: A term weighting approach. *Expert systems with Applications*, 36(1):690–701, 2009.

[16] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[17] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[18] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

[19] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

[20] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. http://www.journalism.org/2014/10/21/political-polarization-media-habits/, 2017. Accessed: 2017-01-22.

[21] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[22] Chris D. Paice. Another stemmer. *SIGIR Forum*, 24(3):56–61, November 1990.

[23] Martin F Porter. Snowball: A language for stemming algorithms, 2001.

[24] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[25] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[26] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[27] Matthew Smith. How left or right-wing are the uk's newspapers? https://yougov.co.uk/news/2017/03/07/how-left-or-right-wing-are-uks-newspapers/, 2017. Accessed: 2017-01-22.

[28] Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM, 2004.

[29] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, pages 1385–1392, 2004.

[30] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[31] Yee Whye Teh, Kenichi Kurihara, and Max Welling. Collapsed variational inference for hdp. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[32] Mark Twain. *The awful German language*. BVK, 1880.

[33] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

[34] www.nltk.org. Natural language toolkit, 2012.