On Domain-specific Topic Modelling Using the Case of a Humanities Journal

Nadja Redzuan[1], Ralf Möller[2], Marcel Gehrke[2] and Tanya Braun[1,*]

[1]*University of Münster, Computer Science Department, Einsteinstr. 62, 48149 Münster, Germany*
[2]*University of Lübeck, Institute of Information Systems, Ratzeburger Allee 160, 23562 Lübeck, Germany*

Abstract

Topic modelling techniques have been an important tool for meaningful information retrieval. One prominent method, latent Dirichlet allocation (LDA), describes documents as distributions over topics and topics as distributions over words. Most applications of LDA focus on sets of tweets, news articles, wikipedia entries, or academic publications covering various topics. How LDA behaves with very domain-specific corpora is less well researched. Therefore, in this article, we apply LDA to infer hidden thematic structures from a corpus of an academic journal concerned with the studies of modern and ancient manuscripts. From this case study, we infer steps specific to dealing with domain-specific corpora.

Keywords

Topic modelling, LDA, Domain-specific corpora

1. Introduction

The rapid digitalisation in the era of Big Data has generated a large influx of textual data. The amount of data created globally per day will reach an estimated 463 exabytes by 2025 [1]. This explosion of data is rendered useless if unstructured data is poorly processed and analysed. Therefore, developments in the area of Natural Language Processing (NLP) have been crucial for handling the increasing volume of textual data [2]. When paired with recent advances in the field of Machine Learning (ML), NLP can greatly facilitate the task of extracting meaningful information from an overwhelming amount of data.

One of the many powerful tools at the intersection of NLP and ML is topic modelling, a computational technique used for information retrieval in domains such as bioinformatics [3], medical science [4], and social networks [5]. Given a large collection of textual data, a topic model can extract hidden topics that best represent the given content. Based on the frequency of observed words in a collection of documents, topic modelling infers underlying themes through patterns of recurring words. Latent Dirichlet allocation (LDA) is a topic modelling technique widely used to infer hidden topics from a large collection of text documents [6]. It has been

3nd Workshop on Humanities-Centred Artificial Intelligence (CHAI)

[*] Corresponding author.

n_redz02@uni-muenster.de (N. Redzuan); moeller@uni-luebeck.de (R. Möller); gehrke@ifis.uni-luebeck.de (M. Gehrke); tanya.braun@uni-muenster.de (T. Braun)
https://www.ifis.uni-luebeck.de/~moeller/ (R. Möller); https://www.ifis.uni-luebeck.de/~gehrke/ (M. Gehrke); https://www.uni-muenster.de/Informatik.AGBraun/team/braun.shtml (T. Braun)
0000-0002-1174-3323 (R. Möller); 0000-0001-9056-7673 (M. Gehrke); 0000-0003-0282-4284 (T. Braun)

CEUR Workshop Proceedings (CEUR-WS.org)

applied to countless applications in the text, image, and video domains have found LDA to be a useful tool for retrieving and analysing information [7, 8].

Although LDA is an established technique to analyse text data in areas such as biomedical science [9] and software engineering [10], scholars in the fields of arts, humanities, and social sciences have yet to fully embrace automated text-analysis techniques [11]. In here lies the motivation for this paper, taking a practical look at topic modelling in areas where data might be more scarce. To that end, we present (a) an optimised LDA topic model for a corpus of a humanities journal and (b) an evaluation regarding how interpretable and how representative the topics are of the collection of documents. We discuss using this setting what topic modelling can offer to the humanities and special steps domain-specific corpora require.

In the following, we briefly recap topic modelling and LDA. Then, we describe our approach and evaluation metrics used to assess a topic model. Afterwards, we report on a topic model learned for a humanities journal. We end with a conclusion.

2. Topic Modelling

In topic modelling, the atomic unit of textual data is a *word*, or otherwise called a *token*. A set of words make up a *document* and a collection of documents make up a *corpus*. Topic modelling adopts the "Bag-of-Words" (BoW) model, in which each document is represented by the number of occurrences of every word in the vocabulary, disregarding the sequential order of words. Suppose a corpus consists of M documents and N words. The BoW model is given by a document-term matrix (DTM) of $M \times N$ dimensions. Based on this BoW model, topic modelling finds patterns of word co-occurrences to infer K hidden topics. *Dimensionality reduction* is the distinctive feature of many topic modelling approaches, whereby a DTM is reduced into document-topic and topic-word matrices of dimensions $M \times K$ and $K \times N$.

LDA is a prototypical topic modelling approach [6]. The key ideas in LDA are that (i) documents are described by probability distributions over topics and (ii) topics are probability distributions over a fixed vocabulary of words. Through the additional assumption that the document-topic and topic-word probability distributions are generated according to Dirichlet priors with hyperparameters α and β, the learning process can be controlled [12].

LDA uses a 3-level hierarchical model, as shown in Fig. 1. The rectangular plates represent replication while arrows denote dependencies. The M plate denotes documents in a corpus and the N plate denotes words in a document. For every document i, there is a multinomial distribution over K topics (θ_i). For every topic k, there is a multinomial distribution over N words (φ_k). Words $w_{i,j}$ are the only *observed* variables whilst the rest are latent variables. Based

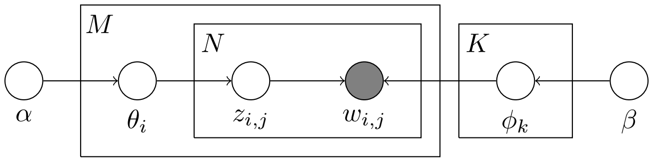

Figure 1: Structure of a graphical topic model as defined in the context of LDA

on the number of topics K, the two prior Dirichlet distributions with parameters $\alpha > 0$ and $\beta > 0$ are set. α influences the per-document topic distribution. A higher α value increases the chance of each topic being assigned to a document, resulting in documents with uniform topic distributions [13]. Likewise, β increases the weight of each word in the topic-word distribution, leading to more uniform word distributions.

The observed and hidden variables along with their dependencies make up the following joint probability distribution for a corpus $\mathcal{D}$:

$$P(w, z, \theta, \varphi; \alpha, \beta) = \prod_{k=1}^{K} P(\varphi_k\,;\,\beta) \prod_{i=1}^{M} P(\theta_i\,;\,\alpha) \prod_{j=1}^{N} P(z_{i,j} \mid \theta_i)\; P(w_{i,j} \mid \varphi, z_{i,j}) \tag{1}$$

which makes LDA a generative model: Drawn from the Dirichlet distribution with parameter α, the latent variable θ_i is the topic distribution for a document i. The latent variable φ_k is the word distribution for topic k drawn from the Dirichlet distribution with parameter β. The variable $z_{i,j}$ is the topic drawn from the multinomial topic distribution of document i for the j'th word. Based on the topic and its per-topic word distribution, a value for variable $w_{i,j}$ is sampled for every j in i.

We approximate a topic model for a corpus by inferring the latent variables θ_i and φ_k from the words in the documents. Given a new document and a learned model, LDA allows for inferring the following distributions based on the observed variables (words): (i) the latent topic distribution θ_i for document i, representing the probability of topics $k = 1, ..., K$ occurring in document i, and (ii) the latent word distribution φ_k for topic k, representing the probability of words $w = 1, ..., N$ belonging to topic k.

3. Learning a Domain-specific Topic Model

This section provides a description of the dataset and methodologies used in the topic modelling task, including the experimental design as well as evaluation metrics used.

3.1. Dataset

The corpus is comprised of a collection of humanities journal volumes, *Manuscript Cultures*, a publication by the "Manuscript Cultures in Asia and Africa" research group of Hamburg University concerned with the study of modern and ancient written artefacts. There are a total of 17 volumes of *Manuscript Cultures* publicly accessible as downloadable PDFs on CMSC's homepage[1]. Each volume contains articles covering various topics, such as manuscript analysis techniques along with calligraphy and writing styles in ancient manuscripts. Volumes 4, 6, 9, 14, 16, and 17 contain no articles as they are exhibition catalogues. Overall, there are 99 articles published between 2008 and 2022 considered as documents for LDA topic modelling.

[1] https://www.csmc.uni-hamburg.de/publications/mc.html

3.2. Experimental Design

The topic modelling process is divided into the following three steps: pre-processing, LDA, and evaluation. The pre-processing module cleans the raw dataset. In the LDA module, the pre-processed texts is transformed into a BoW model, which is then used to train an LDA topic model. The evaluation module involves assessment of the trained topic models.

3.2.1. Pre-processing Module

The objectives of the Text Pre-processing Module are to import, organise, and clean the raw text dataset. This is a crucial first step in the topic modelling process because an unstructured input dataset affects the quality of the trained topic model. Methods such as normalisation, removal of stop-words, lemmatisation, and removal of short words are applied to pre-process the texts. Normalisation of text data involves removing digits and punctuation from the text, and converting the text into lowercase. Stop-words are common words that do not contribute to semantic significance to a text, such as articles, prepositions, and conjunctions. The NLTK python platform provides a list of 179 common English stop-words. Additionally, we extend the list of stop-words with words such as “fig”, “column”, and “ieee”. The goal of lemmatisation is to reduce variation of forms of a word, thus shrinking the size of our input data for topic model training, which benefits from part-of-speech tagging to more accurately identify the root of a word. In addition, we keep only words of length $n > 2$ in our pre-processed dataset as analysing preliminary trained topic models based on pre-processed texts without eliminating single characters and words of length $n = 2$ showed that these short words contribute little to no semantic significance to our corpus.

3.2.2. LDA Module

In the LDA Module, the aim is to train an optimal LDA topic model. The sequence of steps involves tokenising the pre-processed texts and building bigrams, which are pairs of words that appear together with a given minimum frequency, before transforming the tokenised texts into a BoW representation and training a model.

Tokenising and Building Bigrams The first step to prepare the input data as a suitable format for model training is to split the documents into tokens. After tokenising the pre-processed texts, our implementation finds bigrams from the documents. Bigrams are words (or tokens) that commonly appear adjacent to each other. We eliminate bigrams that have a low frequency. Our implementation scores bigrams based on the Normalised Pointwise Mutual Information (NPMI) scoring function. Hence, bigrams have a score between -1 and 1 based on the frequency of the first and second tokens. A higher `threshold` score results in fewer bigrams. The computed bigrams are then added to our list of tokens, so that the BoW we use to train our model is more attuned to the natural language expressions humans use, which often includes pair of words. For example, a topic model that produces a topic with the phrase “fifteenth_century” would be more beneficial for the interpretation of topics, rather than the singular tokens “fifteenth” and “century”.

Transform into Dictionary and Corpus Our implementation removes tokens that appear in less than 20 documents and in more than 70% of our corpus. Doing so produces the best results for training the LDA model in our case. This is followed by transforming our data into a BoW model, where each document is represented by the frequency of every token in our dictionary. The collection of these vectorised documents make up the corpus used as input for training the topic model.

Training the Topic Model For an LDA topic model to extract meaningful and comprehensible topics, it is critical to specify the number of topics we want to obtain at the beginning of the training process. To determine the optimal number of topics K, we train topic models by varying the values of K, ranging from 1 to 40, and evaluate the models using metrics discussed in Section 3.2.3. Through fine-tuning of the parameters during topic model training, an initial base model and a second model are selected for analysis and comparison after multiple training iterations. Furthermore, a final topic model is trained after optimising the hyperparameters α and β. The trained topic models are compared and the results are discussed in Section 4.

3.2.3. Evaluation Metrics

After training multiple LDA topic models, we choose and compare optimal models based on conventional quantitative evaluation metrics, such as perplexity and topic coherence, as well as qualitative metrics based on human interpretation of the topics produced by the model.

Perplexity Perplexity evaluates the predictive quality of a topic model and is calculated based on the log-likelihood of a document. Log-likelihood measures how *likely* the trained model fits the data and is defined as [14]:

$$log\text{-}likelihood(\mathcal{D}) = \sum_{i=1}^{M} \log(P(\,w_i\,)) \tag{2}$$

where $\mathcal{D}$ is a corpus with M documents and w_i are the words contained within the ith document. $P(\,w_i\,)$ denotes the probability of the observed document based on the topics generated from the trained model. A higher likelihood score means the model is better fitted to the trained data. The perplexity is then computed as follows [15]:

$$perplexity(\mathcal{D}) = 2^{-\frac{log\text{-}likelihood(\mathcal{D})}{N}} \tag{3}$$

Lower perplexity scores imply that the trained model is less *perplexed* by the data, hence indicating a better model. As lower perplexity scores do not necessarily correlate to more interpretable topics [16], we also look at topic coherence as an evaluation criterion.

Topic Coherence Topic coherence captures the interpretability of topics and is measured through the degree of semantic similarity between high scoring words in the topic. Röder et al. propose the C_V measure as the best performing coherence model, which is calculated through a pipeline described below [17]: (i) *Segmentation*: From the words with highest probabilities

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
script 0.009	copy 0.008	method 0.017	ink 0.037	gospel 0.018
buddhist 0.005	scribe 0.007	data 0.012	sheet 0.015	codex 0.014
literary 0.004	chapter 0.007	approach 0.010	parchment 0.009	homily 0.013
university 0.004	god 0.007	tool 0.009	papyrus 0.008	fols 0.012
copy 0.004	arabic 0.007	script 0.008	scroll 0.008	liturgical 0.009
catalogue 0.004	index 0.007	digital 0.008	fibre 0.008	greek 0.008
art 0.004	title 0.006	style 0.007	xrf 0.007	fragment 0.007
monastery 0.004	unit 0.006	model 0.007	lead 0.007	church 0.006
literature 0.004	subsection 0.006	user 0.007	light 0.006	john 0.006
print 0.004	table 0.006	task 0.006	pigment 0.006	monastery 0.006

Table 1
Topic-word distribution of TM for the first five topics with $K = 7$. Each column shows the top-10 words and their weights in each topic.

$W = \{w_1, ..., w_n\}$ in a topic k, a set of subset pairs of words (W', W'') with $W', W'' \subseteq W$ is sampled. (ii) *Probability Calculation*: Word occurrence probabilities $P(W', W'')$ based on the corpus are calculated with the *boolean sliding window* method, whereby a sliding window is moved over the text to measure the proximity between words in a document. (iii) *Confirmation Measure*: An indirect confirmation measure is calculated as indication of the relation between subsets of words. The cosine similarity between mixture vectors of words in W' and W'' is the final confirmation score. (iv) *Aggregation*: The arithmetic mean of the calculated confirmation scores is the final coherence value. The C_V measure generates a topic coherence score between 0 and 1. A model with higher topic coherence score extracts topics that make more sense to humans, hence we aim to maximise the topic coherence score when training a topic model.

4. Results

This section provides an in-depth discussion on the evaluation results and decision rationales of the topic modelling process.

4.1. Optimal Model

To find an optimal topic model, we train several topic models and calculate the mean topic coherence scores c. In a first step, the topic models vary in the number of documents processed in each training chunk (`chunksize`), how many times the model is trained on the corpus (`passes`), and the number of topics k. The mean topic coherence scores for `chunksize` $= 10$ are constantly the highest within each iteration of the different `passes` values with $c = 0.492$ for `passes` $= 10$, $c = 0.496$ for `passes` $= 20$, and $c = 0.497$ for `passes` $= 30$. Hence, we set the value for `chunksize` at 10. We vary the value of `passes` $\in \{20, 30\}$. In a second step, we find the optimal `alpha` and `beta` values by training topic models with varying parameter values as follows: • $K \in \{1, ..., 20\}$ • `passes` $\in \{20, 30\}$ • `chunksize` $= 10$ • $\alpha \in \{1/K,\ 0.01,\ 0.1\}$ • $\beta \in \{1/10 \cdot K,\ 0.001,\ 0.01\}$ At the end of the training process, we obtain a model with a topic

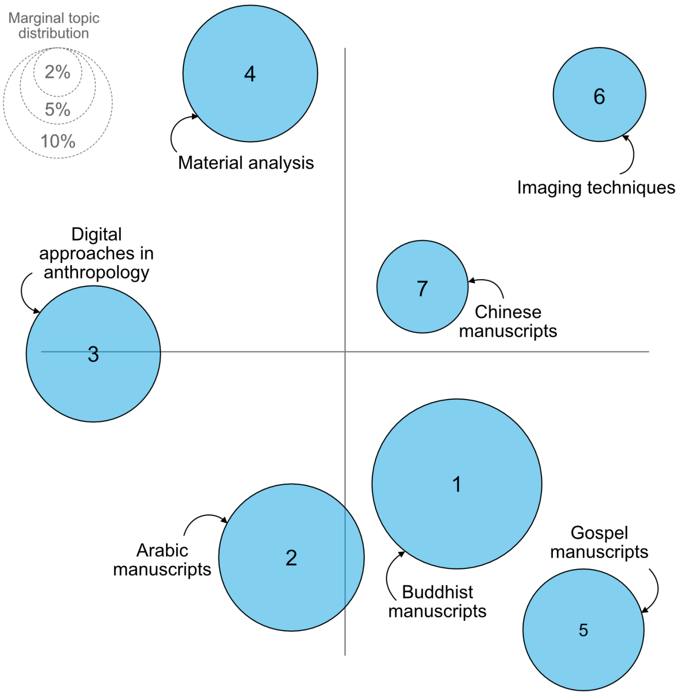

Figure 2: Remodelling of the Inter-topic Distance Maps from `pyLDAvis` for TM, with annotations of inferred topic labels. (The axes do not carry any meaning.)

coherence score of $c = 0.574$ and a perplexity score of $p = 140.492$, whereby the configuration of parameters is ($K = 7$, $\texttt{passes} = 30$, $\texttt{chunksize} = 10$, $\alpha = 0.1$, $\beta = 0.7$).

To get a feeling of the result, let us take a closer look at the optimal topic model, which we refer to as TM. Table 1 lists the top 10 words and their weights for five of the seven topics. Topics 1, 3, 5, and 7 contain similar distinguishing words that make up the topics “Buddhist manuscripts”, “Digital approaches in anthropology”, “Gospel (Georgian) manuscripts”, and “Chinese manuscripts”. Although Topic 4 consists of words such as “xrf”[2] and “light” which could point to an imaging technique, we give it the label “Material analysis” due to the words “ink” and “parchment” having higher weights in the topic. Furthermore, Topic 6 has the bigram “multispectral_image” and the word “light”, providing stronger indication that this topic is related to imaging techniques. Apart from “god” and “arabic”, the words that make up Topic 2 are indistinct, therefore we are left with no option but to label this topic with “Arabic manuscripts”.

The Inter-topic Distance Map (IDM) for TM illustrated in Fig. 2 shows that all topics are distinct from each other, which is an indicator of a good topic model, since the topics do not overlap. We observe that the topics linked to manuscripts are closer together in the bottom right half of the IDM, while the topics “Digital approaches in anthropology”, “Material analysis”, and “Imaging techniques” are dispersed further away.

[2] “xrf” refers to X-Ray Fluorescence, a method for analysing manuscript composition.

4.2. Interpretability

To determine whether the learned model provides interpretable topics and to get closer to form of ground truth, we look at the individual volumes. Figure 3 shows the topic distribution of each of them. To analyse whether they form an accurate representation of the actual topic distribution in our corpus, we manually assign the documents in each volume to a topic label from TM based on human interpretation. We then rate the topic distribution for each volume based on context by referring back to the contents of the volume.

A qualitative rating of "high", "medium", and "low" is given to assess how well the topic distribution compares to the actual corpus. In this case, "high" represents that the topic distribution correlates well with the content and its perceived prevalence in a given volume of *Manuscript Cultures*. "Medium" indicates that the topic distribution is acceptable, but the order of prevalence is debatable. A "low" rating for a topic distribution means that the order of prevalence of the topics and the allocation of the word distribution of the topic itself is questionable in relation to the context of the corpus. Out of the eleven volumes of *Manuscript Cultures*, seven volumes are ranked as "high" (Volumes 1, 3, 7, 8, 11, 13, 15), three are ranked as "medium" (Volumes 2, 10, 18), and one is ranked as "low" (Volume 5). Let us look at the three categories.

Rating "high" Volumes 7, 11, and 15 contain the proceedings of the "Conference on Natural Sciences and Technology in Manuscript Analysis", therefore these volumes contain articles covering themes such as manuscript analysis techniques, exploration of inks and materials in historical manuscripts, and computational methods for the study of manuscripts. Thus, it makes sense that the topics "Digital", "Material", and "Imaging" have high prevalence in these volumes. For another example, we look at Volume 1, which explores Buddhist Sanskrit manuscripts from Tibet and Buddhist scripts from the surrounding areas of Thailand. Furthermore, an article from Volume 1 investigates inventory of tombs of early ancient China. Hence, the high prevalence of the topic "Buddhist" followed by "Chinese" is consistent with the corpus. For the other volumes (3, 8, 13), a comparable interpretation can be found.

Rating "medium" In Volume 2, the topic "Buddhist manuscripts" is listed as the most dominant topic, although two out of five articles are perceived to have more relevance to the topic "Chinese manuscripts", as one article focuses on the study of ancient Chinese manuscripts, while the other article covers Chinese ligatures in Uigur manuscripts from the 13th and 14th century. In contrast, only one article in the volume covers the preservation and documentation of Buddhist manuscripts from monasteries in Laos. The remaining two articles investigate Arabic and Sanskrit manuscripts, respectively. Hence, we infer that a proper descending order of topics is "Chinese", "Buddhist", "Arabic". Similar arguments arise for Volumes 10 and 18.

Rating "low" In Volume 5, one-third of the articles are related to the study of characters, handwriting, terminology, or calligraphy in manuscripts. By intuition, we would like to categorise these articles as "Character and handwriting studies", but since TM does not contain this topic label, a compromise is to assign the label "Chinese manuscripts", since the topic-word distribution for "Chinese" contains the words "character" and "calligraphy". However, this contradicts with the prevalence of the topic "Buddhist" in Volume 5. Two of the articles in Volume

5 study the types of materials used as manuscripts, one for ancient Tamil texts and the other for Tibetan scripts. One article examines the structure of Arabic poems. Thus, the prevalence of the topics “Arabic” and ”Material” is substantiated. The proposed order of significance of topics for Volume 5 based on the context is therefore “Material”, “Chinese”, and “Arabic”.

The Curious Case of “Buddhist” The topic “Buddhist” is the most dominant topic throughout the corpus. It is the topic with the highest distribution in Volumes 1, 2, 3, 5, and 8, having the highest prevalence in Volume 1 with a distribution of 0.803. However, it is apparent that “Buddhist” is commonly listed as a prevalent topic in the topic distributions that are ranked “medium” and “low”. We deduce that the cause of this misclassification is the occurrence of many common words in the topic-word distribution for “Buddhist”. The top words such as “script” and “literary” have a high probability of occurring throughout the corpus due to the context of the subject domain. However, this is not that surprising since the corpus is a collection of rather homogeneous documents which centres around the same specific domain of manuscript studies, therefore the same prevalent words are spread throughout the corpus.

4.3. Discussion

We consider the learning approach and results on a technical level as well as their meaning for humanities research and domain-specific corpora in general.

On a Technical Level In summary, the topics extracted by TM are found to be interpretable, although many common words make up the topic-word contribution. A previous topic model that showed the best performance before we started optimising for different hyperparameters had $K = 11$ topics, which lead to overlapping topics in its IDM, with a cluster of topics labelled “Arabic manuscripts”, “Buddhist manuscripts”, “Gospel Georgian manuscripts”, and “Studies of Chinese characters”, with “Homiletic Greek manuscripts” not far away. That is, the previous model showed even more specialised topics whereas the presented model TM exhibits broader topics. The reason for the more general topics is inferred to be a result of the high β value of 0.7, since a higher β value affects the sparsity of a word in the topic-word distribution, thus resulting in more uniform word probabilities within a topic. An analysis of the topic distribution in each volume of *Manuscript Cultures* reveals that the topic model is a fitting representation of our corpus, although the topics are more general. However, it is important to state that there is space for improvement for the labelling and rating of topics, since it is an interpretive task that heavily benefits from expert knowledge in the domain.

The topic-word distributions of TM contain common words that add to the challenge of interpreting the topics. Taking into account that our corpus heavily focuses on the studies of ancient and modern manuscripts, we can expect that the vocabulary used in the text can be repetitive and dense with specialist terms. Therefore, words such as “script”, “literature”, and “print” in Topic 1 appear extensively throughout the corpus, causing these words to contribute to a higher prevalence in an extracted topic. However, there is also a satisfactory distribution of domain-specific words that could provide context to the user for the interpretation of extracted topics. Hence, it is reasonable to say that TM is an adequate topic model for our corpus.

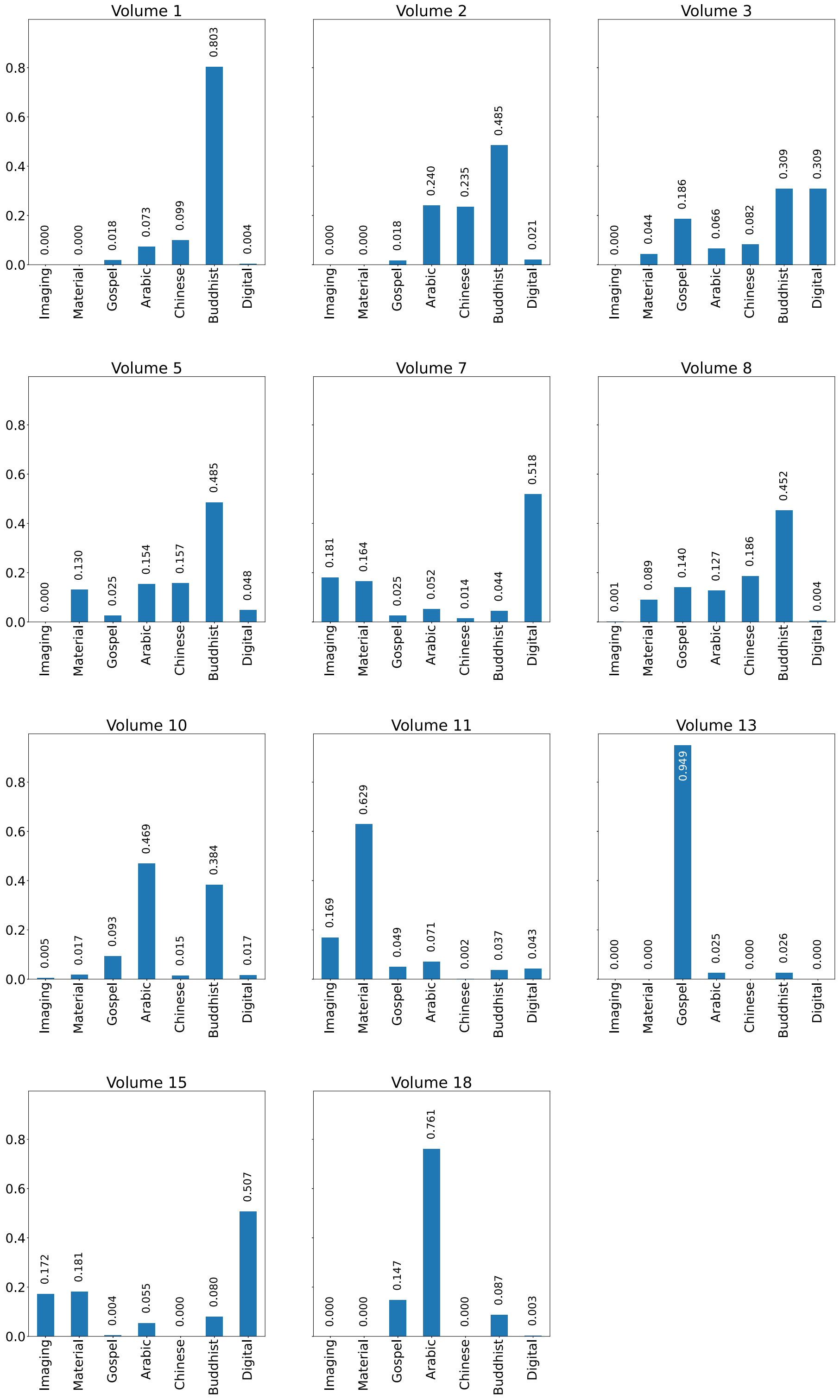

Figure 3: Topic distribution for each volume of *Manuscript Cultures* from TM.

As A Support Tool for Humanities Research As an undersupervised learning technique, topic modelling using LDA can uncover hidden themes as exemplified above: The resulting topic model makes the broad themes covered by the journal explicit, highlighting different techniques, themes, or formats that have influenced the work in the past. By looking at different models, learned with different parameter settings, a topic modelling analysis can reveal broader as well as more specific themes. E.g., as mentioned above, TM generated non-overlapping topics that covered broader thematic patterns, while the topic models with lower coherence scores show overlapping and clustering of topics with very specific words. A given topic model might also provide insight into a new volume or even a new corpus by analysing a document for its prevalent topics or adapting the given model to the new setting, helping to explore the unknown documents without having to read all of them first.

Conjecture Based on the behaviour discussed and the decisions made during preprocessing, we conjecture that when dealing with small, domain-specific corpora, the following observations will be important to consider: (i) The list of stop-words had to be adjusted for the specific setting but even afterwards, domain-specific vocabulary like “script”, “literature”, and “print” as mentioned above still contribute relatively little when interpreting topics. (ii) Choosing an appropriate number of topics K is not easy and can have far-reaching consequences as topics are much more likely to have overlapping vocabulary. (iii) Additionally and as expected, domain knowledge becomes much more important for fine-tuning parameters and interpreting topics.

5. Conclusion

In this paper, we considered the effect of a very domain-specific corpus on topic modelling. Specificially, this paper showed how LDA, a generative and probabilistic topic modelling technique, can be used to discover hidden thematic structures from a corpus consisting of 99 humanities journal articles concerned with the study of manuscripts. We conjecture that domain knowledge and domain-specific settings are necessary to infer interpretable topics. As such, these processes are hard to further automatise, making interdisciplinary work all the more important, especially since domain knowledge as of now is incorporated through hyperparameters that may not appear intuitive from the outset to humanities researchers. Nonetheless, topic modelling also has the potential to provide humanities researchers with a tool for exploring a corpus by its themes without having to read every document in it first.

In the future, we are interested in looking at extensions of LDA to analyse how the methods behave when presented with domain-specific small corpora. These extensions include dynamic topic modelling over time [18] as well as Hierarchical Dirichlet Process [19], which includes finding the right K automatically.

Acknowledgments

The research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany s Excellence Strategy – EXC 2176 ‘Under- standing Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures’, project no. 390893796.

References

[1] P. Taylor, Total data volume worldwide 2010-2025, 2021.

[2] R. Agerri, X. Artola, Z. Beloki, G. Rigau, A. Soroa, Big data for natural language processing: A streaming approach, Knowledge-Based Systems 79 (2015) 36–42.

[3] H. Wang, M. Huang, X. Zhu, Extract interaction detection methods from the biological literature, BMC bioinformatics 10 (2009) 1–13.

[4] Y. Wu, M. Liu, W. J. Zheng, Z. Zhao, H. Xu, Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation, in: Biocomputing 2012, World Scientific, 2012, pp. 422–433.

[5] Y. Cha, J. Cho, Social-network analysis using topic models, in: Proceedings of the 35th ACM SIGIR, 2012, pp. 565–574.

[6] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[7] J.-T. Chien, C.-H. Chueh, Dirichlet class language models for speech recognition, IEEE Transactions on Audio, Speech, and Language Processing 19 (2010) 482–495.

[8] Y. Wang, P. Sabzmeydani, G. Mori, Semi-latent dirichlet allocation: A hierarchical model for human action recognition, in: A. Elgammal, B. Rosenhahn, R. Klette (Eds.), Human Motion – Understanding, Modeling, Capture and Animation, Springer Berlin Heidelberg, 2007, pp. 240–254.

[9] B. He, J. Tang, Y. Ding, H. Wang, Y. Sun, J. H. Shin, B. Chen, G. Moorthy, J. Qiu, P. Desai, et al., Mining relational paths in integrated biomedical data, PLoS One 6 (2011) e27506.

[10] J. C. Campbell, A. Hindle, E. Stroulia, Latent dirichlet allocation: extracting topics from software engineering data, in: The art and science of analyzing software data, Elsevier, 2015, pp. 139–159.

[11] M. Savage, R. Burrows, The coming crisis of empirical sociology, Sociology 41 (2007) 885–899.

[12] H. Wallach, D. Mimno, A. McCallum, Rethinking LDA: Why priors matter, in: Advances in Neural Information Processing Systems, volume 22, Curran Associates, Inc., 2009.

[13] K. Du, Evaluating hyperparameter alpha of LDA topic modeling, 2022.

[14] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (2004) 5228–5235.

[15] M. Hoffman, F. Bach, D. Blei, Online learning for latent dirichlet allocation, Advances in neural information processing systems 23 (2010).

[16] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, D. Blei, Reading tea leaves: How humans interpret topic models, Advances in neural information processing systems 22 (2009).

[17] M. Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in: Proceedings of the eighth ACM international conference on Web search and data mining, 2015, pp. 399–408.

[18] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 113–120.

[19] Y. Teh, M. Jordan, M. Beal, D. Blei, Sharing clusters among related groups: Hierarchical dirichlet processes, Advances in neural information processing systems 17 (2004).