Imperial College
London

COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF LIFE SCIENCE

Comparing Unsupervised and Supervised Latent Dirichlet Allocation (LDA) for Topic Modeling the STEM Research Funding Landscape in UK

Author: Zitong Zhao (CID: 02294840)
Supervisor: Samraat Pawar

Date: August 24, 2023
Word Count: 5992

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at Imperial College London
Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

I declare the data used in this dissertation was provided by Bellotto Trigo, Flavia and Pawar, Samraat.

I declare that I am responsible for the processing, visualisation and analysis of the data.

I would like to extend special gratitude to my supervisor Samraat Pawar, for his dedicated guidance and assistance throughout the research process.

Contents

Abstract

Topic modeling is one of the most powerful techniques in text mining, utilised for data mining, latent data discovery, and identifying relationships between data and text documents. Latent Dirichlet Allocation (LDA) is one of the most popular methods in this field. Additionally, I will introduce a supervised variant of LDA called supervised Latent Dirichlet Allocation (sLDA), which leverages labelled data to classify document topics. The model results for sLda contain estimated parameters, which are used to describe the relationship between labels and topic. The comparison and evaluation of unsupervised and supervised LDA models involve metrics such as semantic coherence, perplexity, and visualisations of high-probability words per topic, providing deeper insights. Through the comparison of LDA and sLDA models, this study enhances people understanding of topic modeling in text analysis applications. The dataset used is the research funding information from the UK Research and Innovation (UKRI). By analysing the distribution of funding within UKRI, it is possible to gain insights into the disciplines at the forefront of development, facilitate interdisciplinary research, and optimise the allocation of funds.

1 Introduction

In today's era of vast electronic text collections, there is an increasing demand for analysing and understanding such large datasets. Natural Language Processing (NLP) is a challenging field of study in computer science that involves information management, semantic mining, and enabling computers to extract meaning from human language processing within text documents [1]. Topic modeling methods are powerful technologies widely applied in natural language processing, enabling the discovery of topics in unstructured documents and semantic mining [2], and has a wide range of applications in various fields [3] [4].

The thesis focuses on comparing the practicality of unsupervised learning and supervised learning in natural language processing. The dataset used is the research funding information from the UK Research and Innovation (UKRI), and topic modeling is employed to classify the distribution of STEM (Science, Technology, Engineering, and Mathematics) disciplines. Subsequently, an analysis is conducted based on the classification results. UKRI is a government-established organisation with the aim of advancing the UK's position as a global leader in science, technology, and socio-economic development. It achieves this by providing support and funding for various activities, including research, innovation, and higher education. This organisation contains numerous approved research proposals, and in this article, I will employ natural language processing to conduct topic analysis and categorise all these proposals. Analysing the distribution of UKRI funding allows us to gain insights into the research directions and significance of various STEM disciplines. It provides valuable information on which disciplines are at the forefront of development. This can promote collaboration and communication between different disciplines, foster interdisciplinary research, and optimise the allocation of funding, ensuring a fair distribution and maximising its impact.

When it comes to classifying a substantial volume of unordered documents into distinct topics, a potent method for accomplishing this task is the utilisation of the Latent Dirichlet Allocation (LDA) model. This model, introduced by Blei et al. ([2], [5]), has been adopted as an innovative approach for document classification. It enables us to reveal the latent themes or subjects present within the documents and discern how words are distributed among these topics. Interest in incorporating prior information into the topic modeling process has grown, with several effective methods developed to model various types of prior information, such as observed labels, latent labels, inter-label correlations, and label frequencies [6]. Additionally, a supervised parameter estimation method based on category and document information can estimate the parameters of LDA according to term weight, while to infer the latent topics that predict the responses, a supervised topic model was developed [7]. One of the advantages of supervised LDA compared to unsupervised LDA is that it can learn topics that align with class labels [8]. Therefore, extending LDA to supervised algorithms is a hot research topic. supervised-LDA is a type of topic model in supervised learning that constrains LDA by defining a one-to-one correspondence between the latent topics of LDA and labels ([9]). Supervised-LDA can directly learn the correspondence between topics and labels. There are also other methods, such as deep learning approaches like CNN, that can be employed for sentiment analysis in topic modeling [10] [11].

In this study, I propose to use a supervised LDA model that leverages label information to guide the learning process of the topic model, leading to a more accurate correspondence between topics and labels. As a result, we expect the labeled LDA model to outperform traditional unsupervised LDA models in classification tasks. Through this research, I aim to validate the utility of machine learning in topic classification and demonstrate the potential of supervised LDA in achieving improved performance in classifying research proposals, providing valuable insights into the landscape of research across various STEM disciplines funded by UKRI.

2 Methodology

2.1 Text Data and Label Preprocessing

The dataset used is the research funding information from the UK Research and Innovation (UKRI). After obtaining a dataset of 85,000 UKRI-funded projects, I initiated the necessary preprocessing steps. This encompassed data cleansing, removing irrelevant characters, special symbols, and non-textual content. Subsequently, I segmented each text into distinct words or tokens, ensuring the text's analysability. Following this, I converted all words to lowercase to address potential issues of word repetition due to varying cases. After removing common stop words, I performed lemmatization on the words, transforming them into their base forms. Eventually, I transformed the tokenized text into a document-term matrix, where each row represents a document, and each column represents a word, thus laying the foundation for subsequent analyses [12].

Each project file contains information about its funding institution, project start and end dates, as well as the total funding amount. I will process this information to derive the labels that I will use for sLDA analysis. Specifically, I will extract the funding institution, funding amount, Subsequently compute the project's duration and the annual funding amount. These processed values will then be used as labels for the sLDA analysis.

After document preprocessing, we trimmed documents with fewer than 20 words and removed documents with missing labels, resulting in a final set of 72,686 documents with clean data. Because a major limitation of supervised learning methods is that they assume the provided labels cover all the classes [13]. After obtaining the cleaned data, I proceed by splitting 75% of the documents as the training set for our utilisation. Initially, I create a corpus from the text data of the training set and then transform this corpus into tokens, essentially dividing the text into individual linguistic units. Subsequently, these tokens are converted into a document-term matrix. Following this process, we proceed to employ this document-term matrix for subsequent model fitting and evaluation assessment.

2.2 Description of the Utilised Model and Analysis Approach

The LDA model operates by using Markov Chain Monte Carlo (MCMC) method for assigning each word in a document to a specific topic based on the topic distribution of that document. By using a fitted LDA model, we can infer the topic structure of an document and make predictions based on this structure. Our goal is to determine the potential topic probability distribution for each document and the potential word probability distribution for each topic. Subsequently, we aim to infer the topics that maximise the posterior probability.
LDA makes two key assumptions:

- Documents are referred to as probability distributions of topics. Each document is a composition of multiple topics, Each topic is associated with a probability. For instance, Document 1 is composed of 65% of Topic 1, 10% of Topic 2, and so on. This signifies that a document might encompass multiple topics simultaneously, with certain topics being more relevant due to higher probabilities.

- Topics are probability distributions of words. Each topic is comprised of a distribution of words. For example, Topic 1 consists of 77% of Word 1, 12% of Word 2, and so forth.

The relationship of probabilities between documents, topics, and words underpins the model [14]. Probability distributions in the LDA model can be summarised as the Dirichlet distribution acts as a prior. It influences the topic distribution in each document and the word distribution in each topic during document generation. Likelihood refers to the probability of generating a document based on known topic and word distributions. The posterior distribution calculates the probability of topic and word distributions given documents and model parameters [15].

During the construction of the LDA model, Gibbs sampling is used, which is a specific instance of the MCMC method for sampling from complex multidimensional probability distributions and generating topic distributions and word assignments. Its fundamental concept involves iteratively sampling variables one by one, given the values of other variables, thereby progressively approximating the desired joint distribution. The model-based classification approach of supervised-LDA requires slightly more additional computation during the testing phase compared to LDA[8].

In the initialisation process, a random topic is assigned to each word within each document. According to LDA, each word is linked to a latent (or hidden) topic. During the iteration, the probability for a word w to belong to a topic t is calculated the product of this two factors: $p(\text{topic } t \mid \text{document } d)$ and $p(\text{word } w \mid \text{topic } t)$ [16].The former represents the proportion of words assigned to topic t within document d, demonstrating the relevance between topic t and the content of document d. The latter signifies the proportion of assignments to topic t from word w across all documents, illustrating the association of word w with topic t across the entire corpus. Based on this conditional probability distribution, a new topic is drawn from the possible topics to update the topic assignment for the current word. After a certain number of iterations, the model converges and attains a stable distribution of topics. Each word within each document is ultimately assigned a definitive topic. LDA endeavors to identify an appropriate distribution of topics and word assignments, maximising the model's ability to generate documents effectively.

In this study, we employ a comprehensive evaluation approach to assess the performance of topic models, incorporating both qualitative and quantitative metrics. For quantitative evaluation, we utilise two widely-used metrics: coherence and perplexity. Coherence assesses semantic consistency within topics, while perplexity measures the model's predictive performance on unseen data. Regarding qualitative evaluation, I employed both human judgment and observation-based assessment. Based on my analysis of the subject categories of UKRI-funded projects, I conducted an evaluation based on the top 20 words of each topic to generate thematic summaries. This approach allowed for a qualitative assessment of the topics and their alignment with the subject matter of the projects.

By combining qualitative and quantitative evaluation methods, this study offers a comprehensive analysis of topic models' performance and usefulness in text analysis applications. I will visualise the results by highlighting the words with the highest probability of belonging to each topic, providing further insights into the identified topics.

2.3 Programming Language and Tools Utilised

I utilised the R programming language to implement LDA and sLDA model. For LDA I used the 'topicmodels' package, which was developed by Grün, Bettina and Hornik, Kurt [17]. LDA operates by probabilistically assigning each word in a document to a specific topic based on the topic distribution of that document.I utilized the 'lda' package to implement sLDA model, which was developed by Jonathan Chang [18]. Unlike traditional LDA, sLDA allows us to incorporate prior knowledge from labeled data to enhance the accuracy of topic classification. The 'lda' package provides functionalities for fitting sLDA models and estimating parameters that describe the relationship between labels and topics. I employed the 'text2vec' package, developed by D. Selivanov [19]. This package offers a range of tools to calculate coherence for topic models, aiding in the assessment of topic quality. It implements various metrics to evaluate topics represented as ordered term lists.

2.4 Coherence

When dealing with a vast dataset of more than 70 thousands UKRI projects, selecting the right number of topics is critical for effective topic modeling. I chose a range of 50 to 500 topics. The dataset covers STEM disciplines and is diverse. Expanding the topic range to 500 captures this diversity well and prevents losing valuable insights due to fewer topics. However, excessive topic numbers can result in redundancy and complexity. After analysing the UKRI disciplines, I've set the upper limit at 500. The chosen topic range balances dataset complexity and model interpretability, leading to accurate and insightful results.

After obtaining the topic classification data generated by both the LDA and sLDA models, I will proceed to assess the coherence of the topics. Topic coherence measures score a single topic by measuring the degree of semantic similarity between high probability words in the topic [20]. When evaluating coherence, several metrics are employed to assess the consistency of topics by machine learning([19]), including logarithmic ratio (Eq. 1), pointwise mutual information (Eq. 2), difference (Eq. 3), and Cosim (cosine of normalized PMI). These indicators help determine the level of coherence by analysing the relationships between individual top words and other top words within a topic. All four of these metrics indicate that larger values correspond to better semantic coherence of the model.

$$\text{logratio} = \log(\varepsilon + T[x,y]) - \log(T[y,y]) \tag{1}$$

$$\text{PMI} = \log_2\left(\frac{T[x,y]}{n_{\text{doc}}} + \varepsilon\right) - \log_2\left(\frac{T[x,x]}{n_{\text{doc}}}\right) - \log_2\left(\frac{T[y,y]}{n_{\text{doc}}}\right) \tag{2}$$

$$\text{Difference} = \frac{T[x,y]}{T[x,x]} - \left(\frac{T[y,y]}{n_{\text{windows}}}\right) \tag{3}$$

In those equations, ε is the smoothing factor. T is the term co-occurrence matrix (which objective is to capture the co-occurrence relationships among words within a text, representing the frequency with which words appear in the same context), and $T[x,y]$ typically refers to the co-occurrence count or frequency of the word ”x” and the word ”y” within a given set of documents, indicating a potential semantic relationship or thematic connection between them. $T[y,y]$ represents the total number of documents used in the calculation of the term co-occurrence matrix. n_{windows} refers to the total number of context windows used for generating the term co-occurrence matrix. A context window is a specified range of tokens surrounding a target token within a text or sequence. In evaluating the coherence of topics by examining the values of these four coherence measurement metrics. The larger the coherence metrics for topics, the more pronounced the overall model's advantage becomes. Higher coherence scores indicate that within specific topics, there is a tighter semantic correlation between words, resulting in stronger mutual coherence among them. Consequently, such a scenario often signifies that the topic model is better able to accurately capture and express the underlying semantic structure hidden within the text data. When assessing the effectiveness and performance of topic models, comparing coherence scores across different topics allows us to draw a more comprehensive and reliable assessment of the overall model's quality.

When evaluating topic models, human judgment through topic consistency evaluation is a common approach. This method involves manually reading and analysing each generated topic from the model to assess its consistency and interpretability.

2.5 Perplexity

To decide on a suitable number of topics, Perplexity-K curve is utilized to decide the optimal K-value [21]. Perplexity is a widely used metric in topic modeling to evaluate the predictive performance of models. It measures how well a model predicts a new set of data given the learned topic distributions. In this study, perplexity was calculated for both the LDA and sLDA models. The perplexity values were obtained by computing the exponential of the negative log-likelihoods of the respective models on the test dataset ([22]).

$$\text{Perplexity} = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{4}$$

M is the number of documents, $\mathbf{w}_d$ represents the words in document d, N_d the total number of words in document d. $p(\mathbf{w}_d)$ is the likelihood of document d, which is calculated from the model's learned topic distributions.

Perplexity provides an effective way to measure how well these models can predict unseen data based on the learned topic distributions. A lower perplexity value indicates better predictive performance and suggests that the model's topic assignments align well with the actual test data.

3 Results

3.1 Exploring Topics and Insights from the LDA Model

In this section, I present a comparative analysis between Latent Dirichlet Allocation (LDA) and Supervised Latent Dirichlet Allocation (sLDA), focusing on their respective performances and capabilities. Both LDA and sLDA are powerful probabilistic topic modeling techniques used for extracting latent structures from text data. Our goal is to highlight the key differences and advantages of these two approaches in various aspects.

In Figure 1, I visualised the topic-word probability matrix of the LDA model which is the posterior distribution of words and documents to topics, and randomly selected a few intuitive topics to examine the generated results, it shows topics are probability distributions of words. Additionally, the word clouds visually represent terms in proportions corresponding to their frequency of occurrence [23]. In each topic's word cloud, it can be observed that the included words exhibit a certain level of correlation. This indicates the capability of the LDA model to categorise text by grouping words with similar semantics. Such clustering aids in identifying commonalities between topics, forming distinctive features, and extracting latent themes from the text.

The size of words indicates their equal probabilities within that specific topic. For example, in Figure 1 (a), these probabilities can infer that this topic is associated with concepts such as females, pregnancy, babies, childbirth, risk, and mothers. This aids in understanding the model's ability to capture and express latent topics within the data.

In Figure 1 (d, e), both visualisations pertain to the field of environmental science and the common keyword in both visualisations is 'ecosystem.' In visualisation (e), the term 'ecosystem' holds significant prominence across the entire topic, signifying its central importance. In contrast, in Figure 1 (d), the term 'ecosystem' (shown at

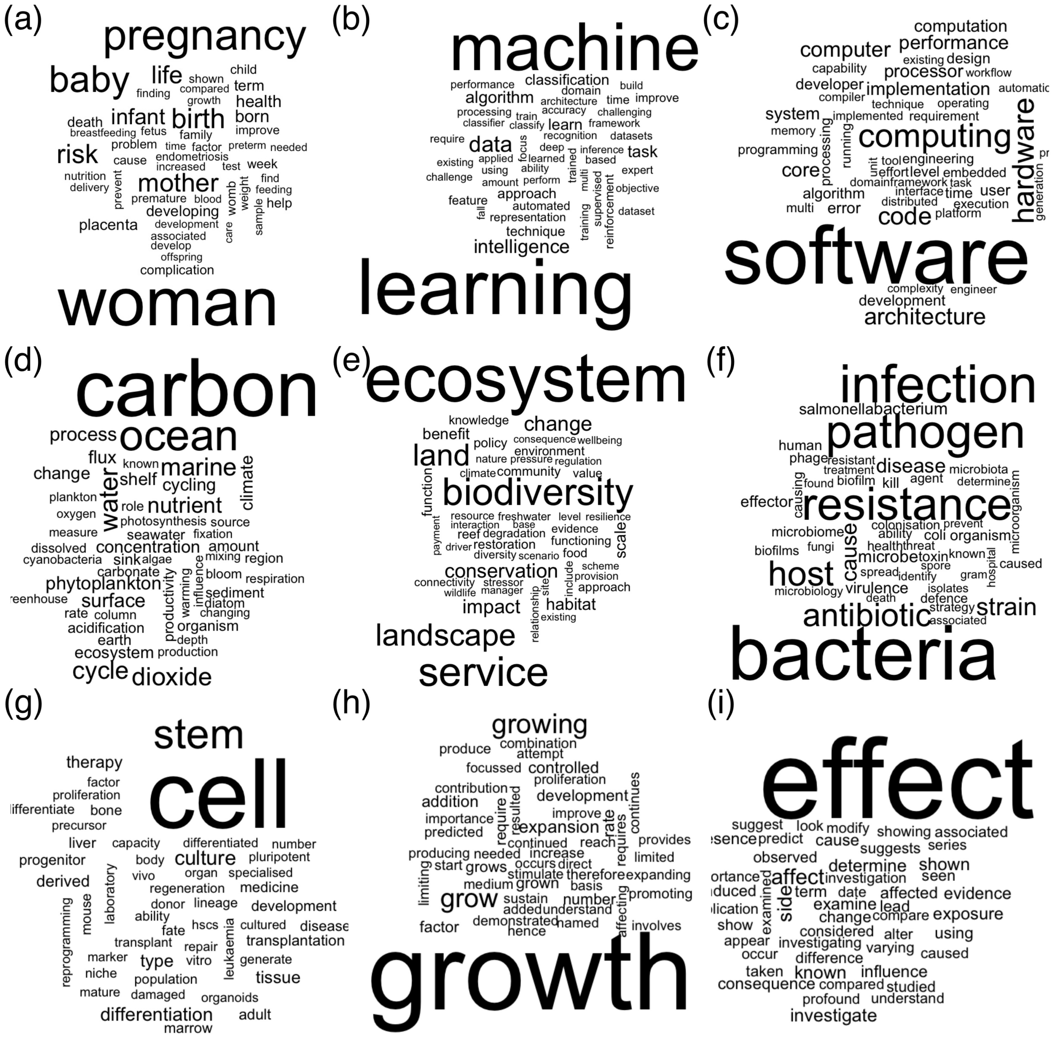

Figure 1: The topic-word probability matrix visualisation of 225-topics LDA model by 'Wordcloud'. These are the topics number 31, 64, 179, 184, 39, 93, 81, 166, and 204 that I selected by their words. The size of words indicates their probabilities within that specific topic.

the bottom) holds a relatively lower probability within the topic. This suggests that the word might possess polysemy or different semantic associations in distinct contexts. This underscores a characteristic of topic models like LDA – they can reveal the complexity of word polysemy and semantic associations across various contexts. This aspect is crucial for extracting deeper insights and meanings from text data.

In Figures (h, i), it can observe topics mainly as ’growth’ and ’effect’. However, these words possess rather general meanings. Consequently, we face difficulty in precisely determining the subject matter of these topics. This uncertainty arises because these words can be applicable in various contexts, such as economy or data-related discussions, potentially relating to the concepts of growth and effect.

This highlights the capabilities of the LDA model in aiding our understanding of inter-topic relationships, identifying commonalities, uncovering latent topics, and addressing the complexities of word polysemy and semantic associations. These insights hold significant value in the realms of text analysis, topic mining, and information extraction.

3.2 Achieving Robust Term Coherence via Optimal Term Co-occurrence Frequency Differences in sLDA with Label Information

The coherence measures the semantic similarity among words within a topic. It evaluates the inherent connections between words within a specific topic context, reflecting the meanings and conceptual associations of these words within that topic. When studying topics within UKRI projects, I explore the degree of correlation across different domains such as environmental science and healthcare, aiming to enhance our understanding of word similarity within topics.

In the context of comparing LDA and sLDA models, the computation of topic coherence becomes particularly crucial. By comparing the topic coherence scores generated by both models, it reveals which model excels in accurately capturing the semantic relationships between words. This process aids in deepening our comprehension of how well the models perform in uncovering hidden topics within the textual data of UKRI projects.

Figure 2 displays the results of four coherence measurement metrics across topics ranging from 50 to 500 for both LDA and sLDA models. The calculation involves the frequency between two terms within the top 20 terms of the topic-word probability matrix of both the LDA and sLDA models. Compared to other metrics, ’difference’ places greater emphasis on the distinctiveness between different terms, highlighting both the frequency disparities among terms and their prominence within specific

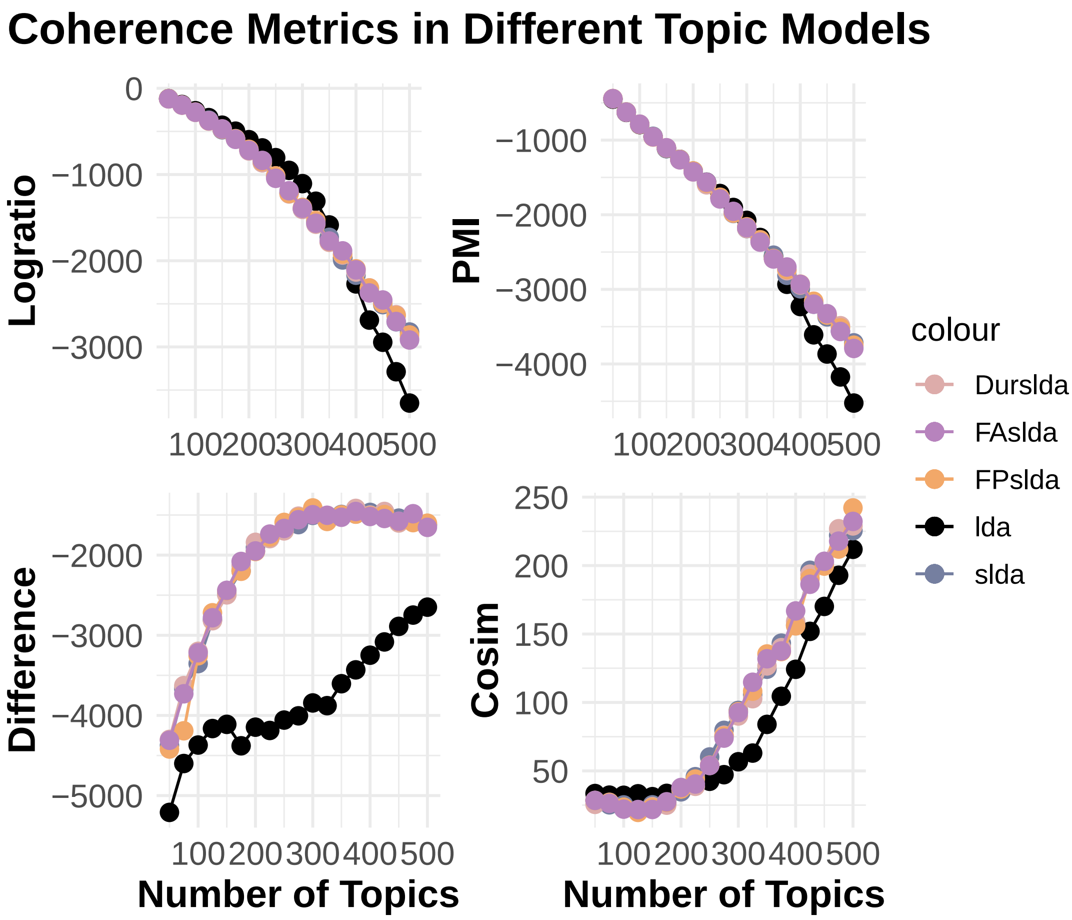

Figure 2: Four coherence measurement metrics across topics ranging from 50 to 500 for both LDA and sLDA models

contexts. Therefore, a larger frequency difference is generally considered more favorable for the model.

In terms of ’logratio’ and ’pmi’ metrics, both LDA and sLDA exhibit similar performances. At smaller numbers of topics, LDA might excel in capturing textual coherence and co-occurrence relationships, thus demonstrating better performance in the early stages. However, as the number of topics increases, sLDA might better adapt to more intricate semantic relationships and inter-topic interactions, resulting in improved performance with larger topic counts. This transition potentially reflects the underlying strengths and adaptability of LDA and sLDA across varying scales of topics.

In terms of the ’difference’ and ’cosim’ metrics, sLDA tends to outperform LDA, which suggests that sLDA holds an advantage in capturing the frequency differences and semantic similarities between terms. Additionally, sLDA exhibits greater proficiency in handling diversity, semantic associations, and topic interactions. By considering label information, sLDA might compute term frequency differences more accurately, leading to higher values when measuring term distinctiveness in the difference metric. Furthermore, the sLDA model introduces label information on top of the LDA foundation, enhancing the semantic representation of terms and enabling the model to establish stronger connections between terms and labels.

3.3 Determine Optimal Topic Counts and Evaluate LDA and sLDA Models Using Perplexity

To make a robust comparison between the LDA and sLDA models and assess the goodness-of-fit and the best topic counts, perplexity calculation serves as a crucial evaluation criterion. Additionally, I sought to identify the optimal number of topics for each model, which further highlights the importance of perplexity analysis in this study.

To calculate perplexity for both models, we utilised the log-likelihood values obtained during model training and then transformed them using the exponential function. The resulting perplexity values provide a quantitative measure of how well the models generalise to new data.

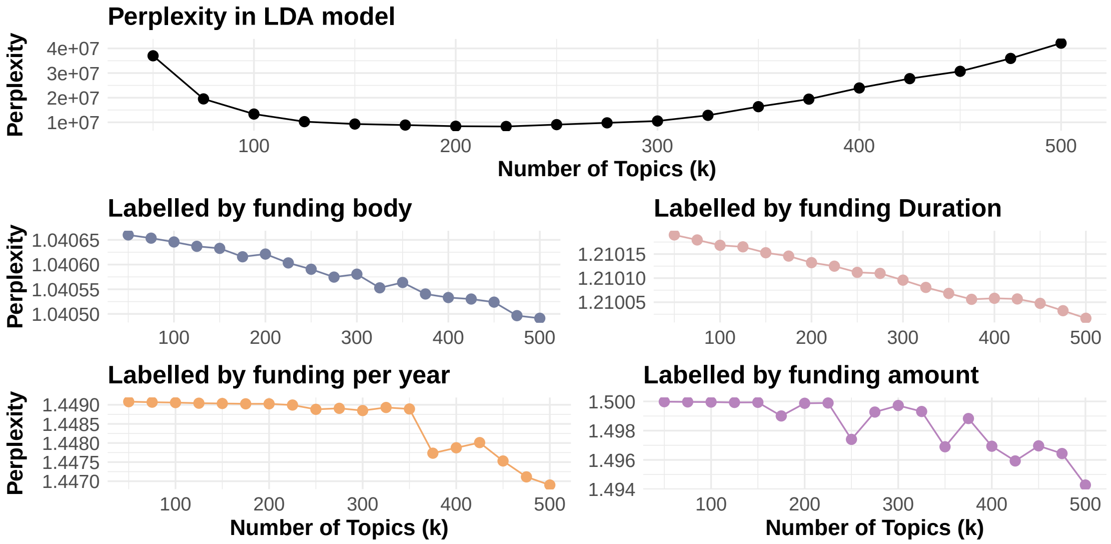

Figure 3: This is the perplexity of the lda model and the slda model in different topics. The slda models have different labels.

The perplexity of the LDA model reached around 2e+07 (in Fig.3), while the perplexity of various labeled sLDA models ranged from 1 to 1.4. This difference could be attributed to the fact that the sLDA model, after incorporating label information, effectively leveraged the underlying structure and latent patterns in the data, leading to superior performance in prediction and lower perplexity.

The LDA model may experience an increase in perplexity due to it overly fit the existing training data, resulting in poorer performance on unseen new text data. Elevated perplexity implies that the model struggles to predict new text data, making it challenging to accurately discern the topics or categories to which the text belongs. Therefore, even though the LDA model might exhibit impressive performance on the training data, its predictive ability could be compromised by its complexity, leading to poorer performance on new text data.

In the LDA model, perplexity is lowest when the number of topics is around 225. This suggests that, on this specific dataset, the LDA model exhibits relatively optimal predictive ability when using approximately 225 topics. Above 225 topics, the perplexity starts to get gradually larger. In traditional LDA models, an increase in the number of topics may lead to overfitting or a rise in model complexity, so that at some point the perplexity starts to rise, creating a clear nadir.

However, it can be seen that in sLDA, perplexity is generally and consistently decreasing as the number of topics increases. The reason is that with a larger number of topics, the sLDA model can become increasingly flexible and better equipped to adapt to complex data distributions and text structures. It can further partition each topic into more specific subcategories, thereby enhancing its ability to represent the text data. This finer-grained classification allows the model to more accurately capture subtle differences between data points, resulting in a gradual reduction of perplexity. Furthermore, the sLDA model with an increasing number of topics might be more effective in classifying texts based on the provided label information. This capability enables the model to more accurately identify differences between various topics, resulting in a gradual decrease in perplexity as the number of topics grows.

After calculating and comparing the perplexity values of the LDA and sLDA models, we found that the sLDA model performs superior in fitting the data. sLDA demonstrates significant performance in prediction, hence possessing superior predictive and categorisation capabilities when dealing with new text data. This results in a notable improvement in perplexity during corresponding evaluations. It can also be stated that the LDA model yields satisfactory outcomes in the analysis of the current dataset. However, its classification performance is evidently less impressive when confronted with entirely new text data.

3.4 Parameter Estimation for Supervised LDA Using Category and Document Information

The sLDA model is additionally capable of estimating the labels associated with each topic it classifies, thereby allowing for parameter estimation in sLDA based on term weights. After training the sLDA model using the ’slda.em’ algorithm in R, the results can be accessed to examine the associations between each topic and corresponding labels. Typically, the output of sLDA model includes Estimate value. This estimates of the sLDA model are obtained through the optimisation of the joint probability of text data and label information. During the training process, the model searches for the optimal parameter settings to maximise the likelihood of the data and determine the probability relationship between each topic and its associated label. These estimated values help us understand the structure and patterns within the text data, providing valuable information for tasks such as text classification, topic analysis, and other natural language processing tasks.

Figure 4 visually presents an analysis outcome of the sLDA model labeled by project duration through data visualisation. I chose a topic number of 75 for a complete presentation and analysis. In the case of a smaller number of topics, I found that the estimated values of sLDA were quite close in this example. The data reveals a distinct pattern where the estimates are predominantly clustered within the range of 800 to 1000. This specific trend is related to the ’technology’ topic, indicating that projects classified under the ’technology’ category exhibit relatively shorter projected duration. The proximity of sLDA’s estimated values could be attributed to the relatively lower complexity of the model under fewer topics, making it more likely to converge to similar estimation results. When the number of topics is relatively low, such as 75, the model’s expressive capability is relatively weak, potentially leading to an insufficient capture of the data’s complexity and subtle differences. As a result, the estimates tend to cluster around a narrower range, which is between 800 and 1000 in this case. With a smaller number of topics, the model has fewer parameters to fit.

In the appendix A, we have included the estimated results of two different sLDA models with varying numbers of topics, specifically 225 and 400. By examining these estimates, we observe an improvement in the model’s performance as the number of topics increases. For instance, when the number of topics is set to 75, the estimates are mainly concentrated between 800 and 1000. However, as increase the number of topics to 225 and then to 400, it can observe a broader range of estimates, including values from 0 to 2500 days. This phenomenon indicates that the model’s performance gradually approaches an optimal state and is closer to achieving the desired classification results.

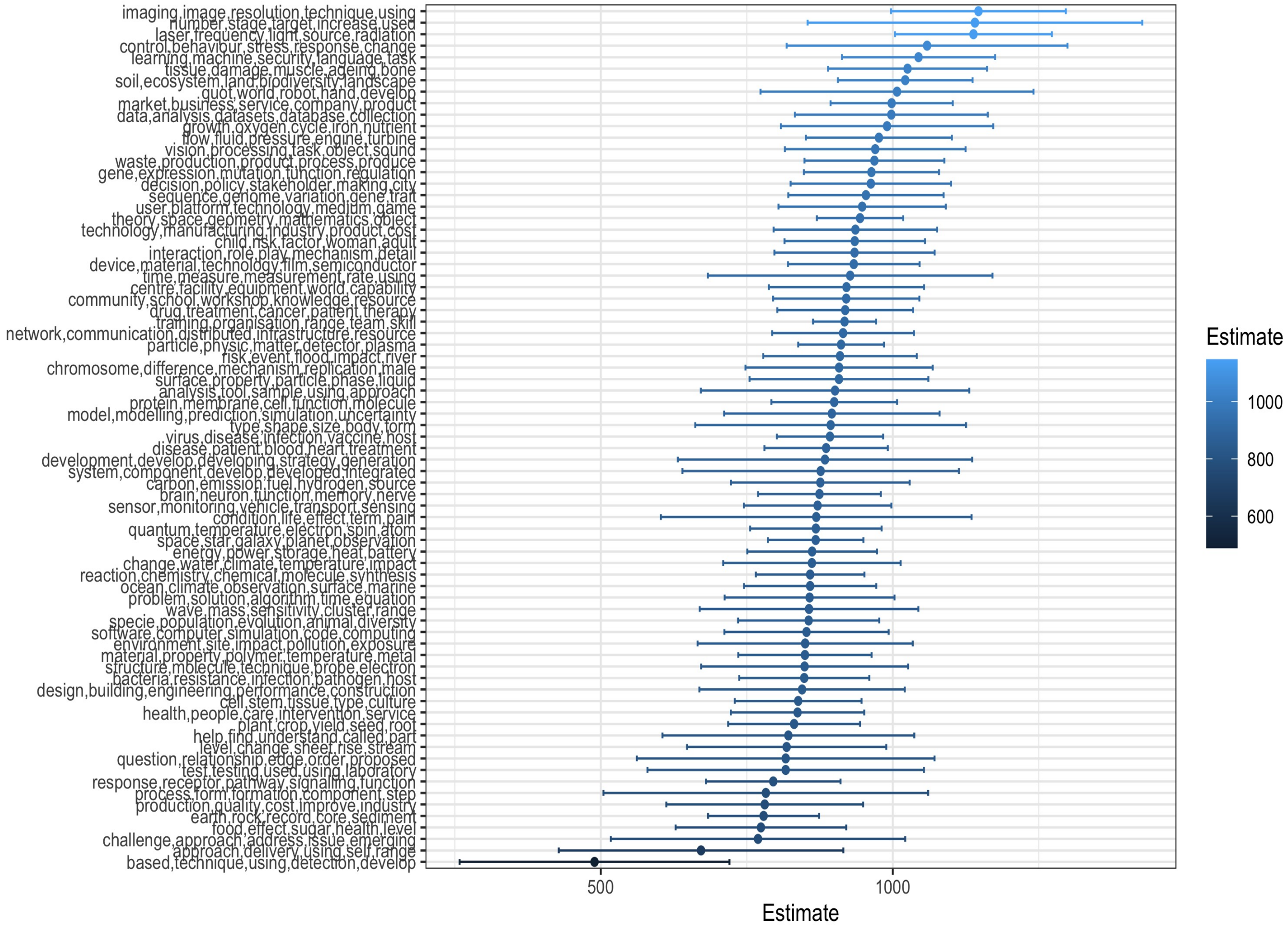

Figure 4: This figure shows the sLDA model labeled by project duration with 75 topics. The horizontal axis represents the coefficients within the sLDA model, which correspond to estimated values for various features. Meanwhile, the vertical axis signifies the different topics generated by the sLDA model. Each point in the figure represents a topic, and the numbers indicate their corresponding Estimate values. Error bars are included, and the upper and lower bounds are derived by adding or subtracting 1.96 times the standard error from the calculated estimate, which portray a confidence interval range of 95%.

With the increase in the number of topics, the complexity of the model also grows, allowing it to better fit the data and capture more underlying structures and hidden patterns within the data. This enhanced complexity and flexibility enable the model to more accurately represent the text data and improve its ability to classify and categorise the documents effectively. As a result, the sLDA model demonstrates improved performance and becomes better equipped to address the complexities present in the dataset.

In the estimated coefficient plot of the sLDA model, if the utilised labels can effectively categorize the topics based on their ranking, then the sLDA model might exhibit improved performance. In the current scenario, there is significant overlap in the estimated values of project duration across different topics. This overlap could stem from the fact that the existing label information is insufficient to distinctly categorise the features related to project duration, leading to overlapping estimated values across various topics. This underscores the significance of labels in the sLDA model. The adequacy of label selection and utilisation can profoundly impact the outcomes and performance of the model. In summary, the sLDA model leverages labeled data to estimate the associations between topics and labels, providing a valuable tool for topic analysis and text classification.

3.5 Topic Consistency by Human Judgements

To enhance the credibility of our analysis in comparing LDA and sLDA, I plan to incorporate human observations and judgments for high-frequency words related to each topic. This approach aims to provide a more dependable assessment. I will personally conduct an assessment of the top twenty most frequent words generated by each model. The selection of 225 topics is based on optimising the LDA model's performance. I selected four similar topics from both the LDA and sLDA models to examine the topic words within each topic (as shown in Table 1 and Table 2). Within these topics, there are identical leading words.

Both LDA and sLDA utilise the Dirichlet distribution as a prior to generate the topic distribution of documents and the word distribution of topics, and employ methods such as Gibbs sampling for model training and inference, contributing to their inherent similarity in the topic generation process. Moreover, while sLDA introduces label information, there are cases where the impact of label information may not be substantial enough to significantly alter the way topics are generated. As a result, some topics in both models may exhibit similarities.

There is an example of viewing the topic words probability. In the topic 'machine learning' in both table 1 and 2, the majority of words are consistent, collectively focusing on aspects such as algorithm and intelligence, so it is not possible to determine which one is better. In the 'ecosystem' topic in LDA, the term 'service' holds a substantial probability, however, 'service' can link to many other topics. Despite this, due to its relatively high probability, the LDA model might inaccurately assign 'service' to the 'ecosystem' topic, potentially causing confusion in topic classification.

Topic	No.	Topic Words Generated by LDA Model
Machine Learning	64	learning, machine, data, intelligence, algorithm, task, learn, classification, approach, technique, feature, improve, automated, reinforcement, time, domain, representation, supervised, trained, performance
Ecosystem	39	ecosystem, service, biodiversity, landscape, land, change, conservation, impact, habitat, benefit, scale, function, policy, restoration, community, reef, environment, food, knowledge, approach

Table 1: I choose some of topic words from the 225-topic LDA output, and summaries the topic name. The value of 'No.' corresponds to the position of the topic within the topic words matrix. The words are sorted in order of their probabilities of composing the topic.

Topic	No.	Topic Words Generated by sLDA Model
Machine Learning	44	learning, machine, algorithm, intelligence, learn, inference, classification, representation, automated, statistic, task, carlo, monte, learned, reinforcement, supervised, setting, trained, segmentation, applied
Ecosystem	221	ecosystem, land, biodiversity, landscape, service, tree, forest, change, vegetation, climate, conservation, scale, habitat, poverty, disturbance, restoration, livelihood, functioning, degradation, agriculture

Table 2: This is the same as Table 1, but it represents the results from the 225-topic sLDA model.

In Figure 5, I have conducted a subject classification using a topic words frequency matrix (for those subject name , I refer to 'chatgpt'). Given the vast scope of the scientific domain, which encompasses numerous subfields, I have undertaken the task of classifying these subjects into smaller, more specific branches.

It can be observed that the overall disciplinary categorisation in both sLDA and LDA is similar within the 225 topics. This similarity might be attributed to the fact that both models utilise consistent priors. I found that in both LDA and sLDA model, also have some mean-less words.

Upon classification of the subject in both sLDA and LDA by myself, I observed that the topics in LDA are often a mixture of two distinct themes. This phenomenon leads to a higher count of subject categories in the LDA inference, most of which are related to biology. Therefore, the frequency of the term 'biology' is notably higher in LDA compared to sLDA. This suggests that in LDA, biological terms are repeatedly categorized across different topics. This multiple classification of biology-related vocabulary across various topics might indicate that the topic divisions in the LDA model are not as precise.

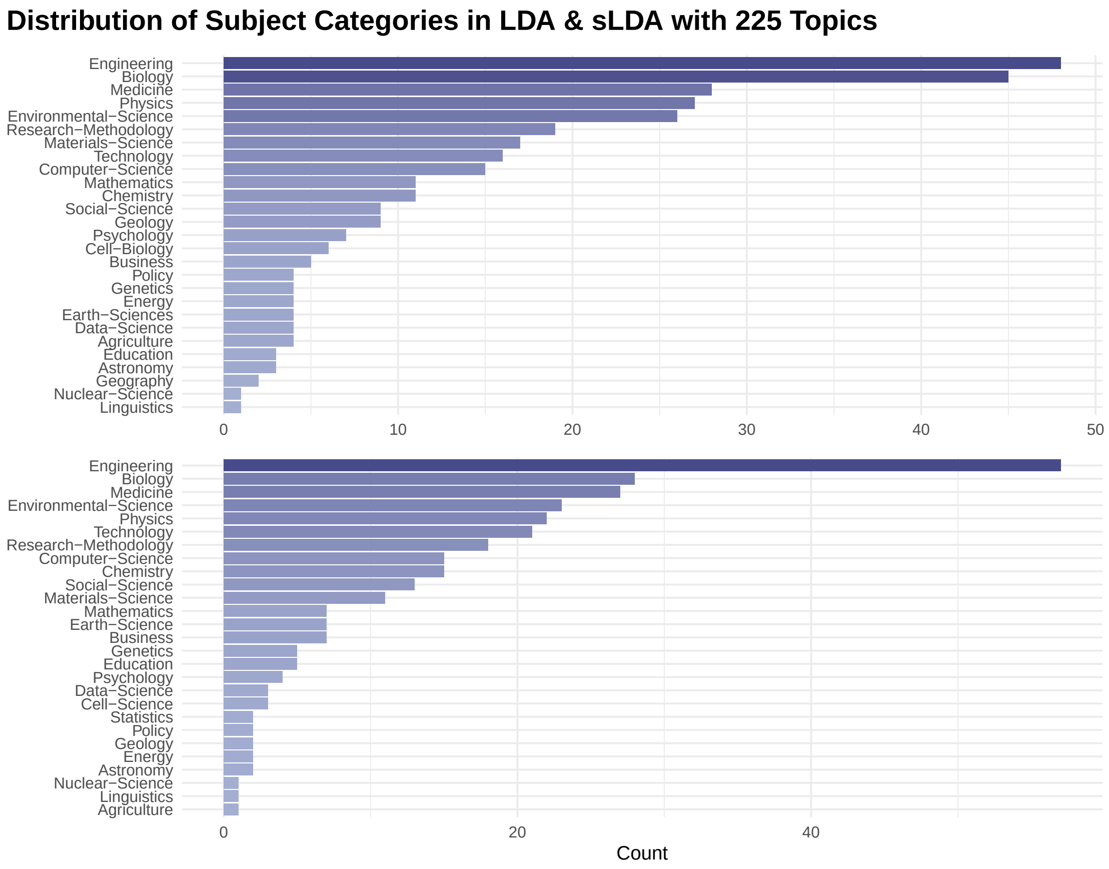

Figure 5: Subject Categories in 225-topic LDA and sLDA. The subject names have been summarised using a subject word frequency matrix. When examining topics resembling those depicted in Figure 1 (h, i), which primarily encompass terms such as ”growth” and ”impact” that lack explicit definitions and pose challenges for classification, I have categorised these topics under the ”Research Methodology” and ”Social Science” domains.

Based on the content of each topic, the majority of UKRI projects cover the field of science, followed by engineering, and then technology and mathematics. This distribution may be attributed to the relatively higher number of projects in the science domain within the UKRI portfolio. This observation potentially reflects the presence of extensive research, exploration, and innovative activities in the realm of science, thereby attracting a larger share of project funding and attention. While engineering, technology, and mathematics also have their share of projects, their representation might be comparatively lower, resulting in a reduced proportion within the overall project landscape. Such a distribution is likely influenced by a combination of factors including funding allocation, demand, research trends.

4 Discussion

The ability of the LDA model to assist us in understanding relationships between topics, identifying commonalities, revealing latent themes, and addressing the complexity of word polysemy and semantic associations holds significant value for text analysis, topic mining, and information extraction.

The LDA model categorises words with similar semantics or thematic associations together, thereby forming distinct characteristics for each topic. Additionally, this clustering of related vocabulary aids in understanding the relationships between different topics and identifying commonalities among them. This further reveals how the LDA model extracts hidden themes from extensive text data, and the generated topic-word frequency matrix allows us to perceive the importance of each word within a topic, thereby revealing underlying content within the topics. Lastly, LDA also highlight the complexity of word polysemy and semantic associations in various contexts, emphasizing that the same vocabulary might hold different meanings across different topics. In conclusion, these observations offer an effective approach for comprehending and analysing the semantic and thematic relationships within textual data.

In the comparison of term coherence, concerning the term co-occurrence frequency metrix, LDA and sLDA exhibit similarities, but they may have different advantages under various topic scales. In terms of the metric reflecting the frequency differences between terms' co-occurrences, sLDA outperforms LDA, indicating that sLDA holds an advantage in capturing frequency disparities and semantic similarity among terms. sLDA also demonstrates greater efficacy in handling diversity, semantic relationships, and topic interactions. By incorporating label information, sLDA calculates larger term co-occurrence frequency differences, leading to higher values in assessing term coherence.

Similar topics and coherence matrices are generated in the sLDA model. This could be attributed to the relatively limited influence of label constraints on topic generation within the sLDA model, resulting in less distinct differences between topics associated with different labels. As a consequence, the coherence matrix exhibits similarities. If the labels used can clearly categorise the topics according to their ranks, this would lead to an improved performance of the sLDA model. Currently, there is significant topic overlap in the estimated values related to project duration. If more precise and distinct labels could guide the model, resulting in a clearer division of topics, it would help reduce topic overlap and enhance the model's performance and accuracy.

When calculating perplexity, the LDA model has a distinct optimal number of topics, while the sLDA model improves as the number of topics increases. This is because the sLDA model is designed to obtain more accurate estimates as the number of topics increases, which in turn exhibits better topic model classification capabilities. We can observe a significant difference in perplexity between the LDA and sLDA models, which could be attributed to the incorporation of label information in sLDA. The integrating of label information in sLDA through weighted parameters may control the influence of labels on the model, and the inclusion of estimated values could enhance predictive capabilities. Different weight settings might lead to varying degrees of label integration into the model. Consequently, sLDA exhibits superior predictive and categorisation abilities when faced with new text data, thereby affecting perplexity. The LDA model is likely to be more complex, leading to high levels of confusion and resulting in a diminished ability to predict new text. The sLDA model is capable of acquiring more prior knowledge about the data structure, enhancing clustering effectiveness and promoting tighter grouping of texts within the same category. This enhancement further contributes to a reduction in perplexity. These findings hold significant implications for text analysis and topic modeling domains, offering potential directions and explanations for further research endeavors.

The estimation values of the sLDA model are obtained by optimising the joint probability of text data and label information. During the training process, the model seeks the optimal parameter settings to maximise the likelihood of the data, and it determines the probability relationships between each topic and its associated labels. This assists in comprehending the structure and patterns within the text data, thereby leading to a better understanding of the content of the UKRI project. The sLDA model not only categorizes each topic but also generates estimation values to enhance the model's performance and predictive capabilities, facilitating label predictions for subsequent projects within the UKRI by leveraging the estimated values.

Through human judgment of the two models, it can be inferred that the LDA model may have some limitations in terms of topic classification accuracy, including potential topic mixing and some irrelevant words. On the other hand, the sLDA model, after incorporating label information, demonstrates an improved ability to address these issues and enhance its performance. While human judgment might not easily discern such differences, machine learning methods can offer more objective and accurate outcomes in performance evaluation. This underscores the significance of employing machine learning techniques for model performance assessment, facilitating a more comprehensive understanding of the models' performance in handling complex data.

5 Appendix

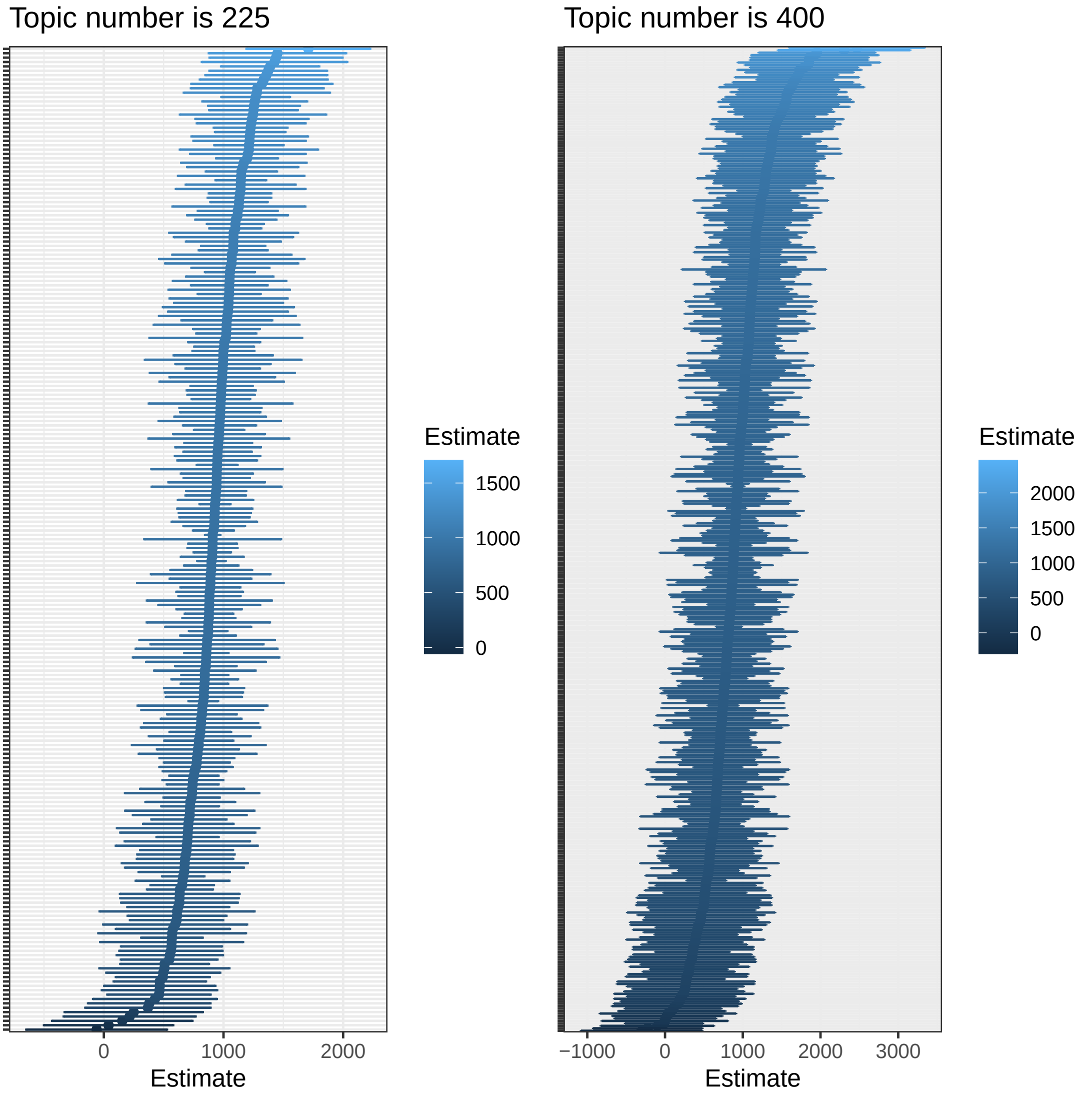

Figure 6: This figure shows the sLDA model labeled by funding body with 225, 400 topic

6 Data and Code Availability

The data and code used in this study are available on [https://github.com/zitong27/Master-Project].

References

[1] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019. pages

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

[3] Arun Varghese, Michelle Cawley, and Tao Hong. Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts. *Environment Systems and Decisions*, 38(3):398–414, 2018. pages

[4] George Papadatos, Gerard JP van Westen, Samuel Croset, Rita Santos, Simone Trubian, and John P. Overington. A document classifier for medicinal chemistry publications trained on the chembl corpus. *Journal of Cheminformatics*, 6(1):40, 2014. pages

[5] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012. pages

[6] Wei Wang, Bing Guo, Yan Shen, Han Yang, Yaosen Chen, and Xinhua Suo. Twin labeled lda: a supervised topic model for document classification. *Applied Intelligence*, 50(12):4602–4615, 2020. pages

[7] Liu Zhenyan, Meng Dan, Wang Weiping, and Zhang Chunxia. A supervised parameter estimation method of lda. In Reynold Cheng, Bin Cui, Zhenjie Zhang, Ruichu Cai, and Jia Xu, editors, *Web Technologies and Applications*, pages 401–410, Cham, 2015. Springer International Publishing. pages

[8] Balaji Lakshminarayanan and Raviv Raich. Inference in supervised latent dirichlet allocation. In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2011. pages

[9] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256, 2009. pages

[10] Ammar Mohammed and Rania Kora. Deep learning approaches for arabic sentiment analysis. *Social Network Analysis and Mining*, 9(1):52, 2019. pages

[11] Abubakr H. Ombabi, Wael Ouarda, and Adel M. Alimi. Deep learning cnn–lstm framework for arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10(1):53, 2020. pages

[12] Ekaterina Kochmar. *Getting Started with Natural Language Processing*. ISBN 9781617296765. Manning Publications, September 2022. pages

[13] Dongyeop Kang, Youngja Park, and Suresh N. Chari. Hetero-labeled lda: A partially supervised topic model with heterogeneous labels. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 640–655, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. pages

[14] Neha Seth. Topic modeling and latent dirichlet allocation (lda) using gensim and sklearn. `https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-skl`
June 2021. pages

[15] Haaya Naushan. Topic modeling with latent dirichlet allocation. `https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8`, December 2020. pages

[16] Ria Kulshrestha. A beginner’s guide to latent dirichlet allocation(lda). `https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2`, July 2019. pages

[17] Bettina Grün and Kurt Hornik. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011. pages

[18] Jonathan Chang. lda: Collapsed gibbs sampling methods for topic models. `https://cran.r-project.org/web/packages/lda/index.html`, November 2015. pages

[19] Dmitriy Selivanov. text2vec: Coherence metrics for topic models. https://search.r-project.org/CRAN/refmans/text2vec/html/coherence.html, 2016. pages

[20] Shashank Kapadia. Evaluate topic models: Latent dirichlet allocation (lda). `https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0`, August 2019. pages

[21] Ling Huang, Jinyu Ma, and Chunling Chen. Topic detection from microblogs using t-lda and perplexity. In *2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW)*, pages 71–77, 2017. pages

[22] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009. pages

[23] Wouter van Atteveldt. Fitting lda models in r. `https://i.amcat.nl/lda/2_lda.html`, February 2018. pages