DALARNA UNIVERSITY

Degree Project

Level: Master's

An investigation of the necessity of incorporating contextual data in topic modeling

Author: Anusha Vaddavalli (h19anuva@du.se), Vishnu Koneswaran (h19visko@du.se)
Supervisor: Moudud Alam (maa@du.se)
Examiner: Siril Yella (sye@du.se)
Subject/main field of study: Microdata Analysis
Course code: MI4001
Credits: 30 ECTS
Date of examination: 20/01/2022

Abstract

This study explores the importance of contextual information in implementing topic modeling algorithms through the analysis of Swedish parliamentary debates. The results from the data analysis enable us to assess the effectiveness of the Swedish parliamentary system by testing the hypothesis that, in an effective parliamentary system, the contents of the parliamentary speeches were governed by the topic of discussion, not by the speech giver's identity. The data is obtained from the Språkbanken text tool provided by the Swedish language technology research unit at Gothenburg University. The text in the tool is enriched with semantics and enhanced with metadata by researchers. Four hundred ninety-one documents of five subjective debates from 2014 to 2018 financial years were taken for the analysis. Considering speech is often the primary way for expressing and communicating political ideas and views, statistical analyses of political text get a long tradition in political science. (Grimmer & Stewart, 2013). To connect our analysis with the ministries of the Swedish government, we further categorized the debate documents to examine the framing of political speeches based on the speech giver's position and party affiliation. This study implemented Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM) algorithms with an unsupervised learning approach. The data was then further categorized, the models rerun, and the topic predictions visualized to evaluate the models regarding the subject of the debate. A simulation study was also conducted to evaluate the performance and uncertainties of the models. Bootstrap sampling was implemented with two hundred replications on the LDA and STM models. Even though the topic loadings were lower in many cases, for both models, the mean confidence intervals of the parameter estimation demonstrated that the activities loaded as a single topic very consistently. As such, the topic modeling results allow us to assess the effectiveness of the parliamentary system through hypothesis testing, where the null hypothesis could not be rejected, i.e. there was insufficient evidence to reject the assertion that the topic of discussion governed the Swedish parliamentary debates

Keywords: Topic modeling, LDA, political science, STM, sampling distributions, hypothesis testing, bootstrap, mean confidence interval

Acknowledgments

We want to express our gratitude to our supervisor, Moudud Alam, for providing us with support and guidance at each stage of the thesis work. We would also like to thank our tutors during the Data Science course for their valuable collaboration throughout the academic tenure. In addition, we would like to thank our family and friends for their invaluable support and discussions in aid of accomplishing this thesis.

Anusha Vaddavalli and Vishnu Koneshwaran
January 2022

Abbreviations

NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
BoW	Bag of Words
STM	Structural Topic Model
LSA	Latent Semantic Analysis
DTM	Document Term Matrix
CTM	Correlated Topic Model
DMR	Dirichlet Multinomial Model
SAGE	Sparse Additive Generative Model
POS	Parts of Speech
EU	European Union
NLTK	Natural Language Toolkit
TF-IDF	Term Frequency-Inverse Document Frequency
SD	Sweden Democrats
KD	Christian Democrats

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Topic modeling is a natural language processing (NLP) technique to label topics by collecting a group of similar words from a collection of documents (Blei & Lafferty, 2009). Labelling assists in classifying, exploring and summarizing the large text that would be difficult through human annotation (Blei, 2012). In the context of text mining, the implementation of topic modeling algorithms is one of the thriving areas to identify the abstract patterns of text (Jelodar, et al., 2019). These algorithms are probabilistic statistical techniques to cluster similar terms in a collection of discrete documents (Blei & Lafferty, 2009). Identifying the topics is beneficial with much of the information available on platforms (Blei & Lafferty, 2009). Applying topic modeling on a high-quality document enables the end-user to understand the content effectively (Blei, 2012).

Topic modeling in Social science research is one of the emerging fields for extracting latent topics, and estimating political variables by using document meta-data (Grimmer & Stewart, 2013). The documentation of Swedish Parliamentary debates has been made openly available from 2010 through the “Riksdagens öppna data” website (Eide, 2020). The researchers enhanced the documents of the Swedish parliamentary debates from 1994 to 2018 (Eide, 2020). The text is annotated with semantics and enhanced with the required fields of metadata (Eide, 2020). Utilizing Parliamentary debates for topic modeling has the distinct advantage that topic activities were well-defined and can be used to evaluate the result of the fitted model by mapping topic labels with documents. In this thesis, we implemented topic modeling algorithms through unsupervised learning by utilizing the advantages of the improved corpus.

Latent Dirichlet Allocation (LDA) by Blei et al. (2003) was a popular topic model for classifying data (Tong & Zhang, 2016, May). Based on the importance of topic modeling with LDA, the current implementation aims to discover topics by analyzing Swedish Parliamentary debates using the LDA algorithm. One of the limitations of LDA is that it assumes words of the documents are stemmed from a mixture of topics without considering drawing inferences on topic proportions (Blei & Lafferty, 2007). The second limitation of LDA is the Bag of Words (BoW) approach, which does not consider the order of words and metadata of the documents (Jelodar, et al., 2019).

Additionally, the Structural topic model (STM) by Roberts et al., (2019) was considered to investigate whether including the documents' metadata can improve recognizing the topics and estimating the relationship of contextual information. According to Roberts et al., (2019), STM successfully models the topics in various domains such as social media, political surveys, international newspapers, and much more.

We conducted a simulation study with the bootstrap method to investigate whether the topic modeling results could be used to assess as a confirmatory tool.

1.1 Background

To understand politics, we should first understand what political actors are saying, and writing this acknowledges the importance of language in the study of political science (Grimmer & Stewart, 2013). Political scientists were using available data sources and statistical methods to communicate political ideas, which include inferring actors' policy positions (Slapin & Proksch, 2008), identifying topics (Hopkins & King, 2010), examining the role of language ambiguity (Campbell, 1983), and estimate the degree of concreteness in political statements (Eichorst & Lin, 2019).

The LDA model was implemented by Grimmer (2010) and further extended to capture the political agenda. The LDA model was expanded by Quinn et al. (2010) inserted a dynamic nature to analyze speech over time gradually. The evolution of the European Parliament's agenda was studied by Greene & Cross (2015) using non-negative matrix factorization.

Analysis of the Swedish parliamentary debates by Magnusson et al. (2018) used party-specific tokens to determine the framing of speeches; with this motivation, this thesis attempts to analyze the relationship between speech properties and draw inferences by including metadata of the debates. An STM model was implemented by incorporating contextual information.

According to Greene & Cross (2015, June), Grimmer (2010), Quinn et al. (2010), and Baumer et al. (2017), topic modeling has so far been designed as an exploratory tool for data classification. There is a research gap with concerning the issue of whether these models can also be used for confirmatory analysis.

1.2 Objectives

This project highlights the importance of using contextual information, in topic modeling via a simulation study. It also aims to compare the performance of STM that includes document level covariate information (that incorporates contextual data and improves inference of topical content) and LDA without contextual data and covariate information, in classifying Swedish parliamentary debates. In short, this thesis attempts to address the following specific objectives.

1. Analyzing Swedish parliamentary debates to assess whether the use of metadata can aid in extracting more meaningful topics than what LDA does.

2. Conducting a simulation study to assess the uncertainties in the topic identification performance of topic modeling.

3. The results from the data analysis enable us to assess the effectiveness of the Swedish parliamentary system through testing the hypothesis that, in an effective parliamentary system, the contents of the parliamentary speeches are governed by the topic of discussion, not by the speech giver’s identity.

1.3 Dissertation of the Report

The rest of this report is organized as follows: In chapter 2, we reviewed the methods, loading of data, data extraction, and pre-processing procedures, as well as the technical aspects of LDA, STM, and Bootstrap sampling. Visualization plots follow this section; chapter 3 explains the results of the fitted models. Chapter 4 was dedicated to discussing the model's findings and performance, the conclusion of the study, and scope to the future work. References were then presented. Lastly, we added the Appendix section, containing Appendix A with table A listing Sweden Government ministries, Appendix B with table B listing Parliamentary Debate Spokesperson Roles, and Appendix C with tables C1 and C2 listing Quality metrics of the fitted models for both LDA and STM.

1.4 Scope of the Study

This thesis aims to examine the impact of using metadata to construct topic modeling algorithms for parliamentary debates. The data used for the analysis were five subjective Swedish Parliamentary debates (see Table 1) that were publicly accessible from 2014 to 2018 via “Riksdagens öppna data website” (RIKSDAG, 2021). The collected speech text consists of 491 documents with 7,083,877 tokens, an average of 14427 tokens per document.

In order to analyze the text and evaluate the results according to parliamentary ministries, data are discriminated by inspired theory from Meguid (2005). As the debates cover various of topics ranging from climate to economics of the industry, the debates were further labelled by considering the responsible ministry during the debate. The resultant corpus contained 16 categories of ministry affairs with 1632 debate documents (see Table 2).

The BoW approach was used to obtain tokens during the LDA implementation in this analysis. Additionally, the STM implementation investigated the properties of party affiliation and the speech giver’s position. We compared the performance of the models with the goodness of fit metric log-likelihood.

Bootstrap sampling was performed on both LDA and STM algorithms and gamma posterior probabilities were calculated. On these posterior probabilities, confidence intervals were calculated, and uncertainties were examined.

Chapter 2

Methodology

This chapter explains the introduction to the topic modeling algorithms used for the analysis, data in detail, data preprocessing steps, and the implementation steps of the models and algorithms used for the study.

The Topic Modeling era started with implementing a probabilistic Latent Semantic Analysis(LSA) algorithm which assumes word frequencies are independent of the document (Hofmann, 1999, August). Later LDA was developed that learns latent topics of the documents by identifying the similarities with a distribution of words (Blei, et al., 2003). This statistical topic model largely depends on the BoW's assumption where word frequencies are quantified as a document-term matrix(DTM) (Blei, et al., 2003). This approach gains the advantage of reducing the computation expense by ignoring the word order. Unlike the micro-level datasets where semantic information is needed to identify topics, a corpus consists of a large set of documents gained by constructing the foundation through the distribution of interconnected documents (Wesslen, 2018). The current study was focused on implementing LDA as the corpus contains 491 documents with 5.8M tokens.

At the same time, the standard topic model could not draw causal inference and multi-modality (Wesslen, 2018). A correlated topic model(CTM) extends LDA which correlates topic proportions with a logistic normal distribution (Blei & Lafferty, 2007). A fast variational framework has been introduced for CTM to compute posterior probabilities that gain predictive power over LDA (Wesslen, 2018). This process, on the one hand, supports an efficient base framework. On the other hand, the model cannot include metadata covariate information (Roberts, et al., 2019).

The implementation steps for the Dirichlet-multinomial model (DMR) include document metadata (Mimno & McCallum, 2008). The inclusion of covariate information to the DMR model affects the topic proportions in the outcome. However, word distributions with DMR are not feasible because they increase computation complexity (Wesslen, 2018). Furthermore, the model does not estimate covariate relationships between topic proportions (Wesslen, 2018). Another framework, which solves the above issues, is introduced with Sparse Additive Generative Model (SAGE) that utilizes log-frequency for distributions (Eisenstein, et al., 2011). Furthermore, the model does not provide the estimation of covariate relationships between topic proportions (Wesslen, 2018).

STM was introduced by augmenting the factors of the three models: CTM, DMR and SAGE (Wesslen, 2018). The major advantage of the STM model is that framework for facilitating hypothesis testing for the impact of document metadata that affects which and how the topics vary by document (Roberts, et al., 2019). Additionally, the model introduces enhancements to the computational methods to make the model feasible

for modeling and methods for model evaluation, interpretation, and handling multi-modality (Wesslen, 2018).

STM was considered for implementation by including document-level information of debates such as party affiliation and role of the speaker to estimate effects of topic proportions.

Roberts et al. (2016) analyzed the estimation benefits of STM relative to LDA and STMs sub-component models like CTM, SAGE, and DMR. They found that for metadata-generated topic processes, STM outperforms LDA in covariate inference and out-of-sample prediction. The current paper focused on conducting a simulation study to evaluate the performance of LDA and STM and estimate the uncertainties with hypothesis testing.

According to Harden (2011), a Bootstrap sampling technique outperformed for validating clustered observations compared to robust cluster standard error(RCSE) approach and Monte Carlo simulations through Ordinary least squares(OLS) regression models. A custom bootstrap model was implemented through 200 random samples on both LDA and STM models, and uncertainties were assessed with mean confidence interval statistics. LDA, STM, and Bootstrap algorithms were all executed using R libraries.

2.1 Loading of Data

The dataset is the Swedish Parliamentary debates available from 2010 through the Riksdagens öppna data website (Eide, 2020). Dataset is imported from Språkbanken text tool, a Swedish language technology research unit at the University of Gothenburg. The corpus in the tool is available in XML format by splitting into words in the Swedish language. The entirety of the data from 1993 to 2018 is available as a single 32GB XML file and downloaded from (SpråkbankenText, 2018). In order to access the required data XiMpLe XML Editor (Novak, 2018-2021) is used to split the 32GB data into smaller chunks of files. The resulting files of five distinct debates between the period 2014 to 2018 contain 55 XML documents each with 100MB of data.

At first, the required XML parent-child nodes of text, sentence, and word and concerning attributes of the debates were manually identified. Python programming is used to extract the required fields of the speech text through XML ElemenTtree API (Python Documentation, 2001-2022).

The text contains 0.8% of end-line hyphenations, cleaned using a rule-based approach to adopt Swedish compounding rules (Eide, 2020). The words of the debate are annotated with linguistic information, parts of speech(POS) tagging (Kumawat & Jain, 2015), and augmented with semantics (Eide, 2020) using the Sprav annotation pipeline (Borin, et al., 2016, November).

The content of the corpus is explained in table 1. The property Chamber activity explains the context of the debate and the number of documents is a specific number of debates that happened during the period from 2014 to 2018.

Table 1: Types of Parliamentary Debates

Chamber Activity	Number of Documents	Financial Period
Interpellation Debate	249	2014-2018
Debates on Parliamentary Business	201	2014-2018
Special Debate	29	2014-2018
Foreign Policy Debate	4	2014-2018
Budget Debate	8	2014-2018

2.1.1 Foreign Policy Debate

The Sweden government is responsible for making foreign policy through the committee on Foreign Affairs. The debates in the parliament follows to shape the policy and make decisions on the international operations. The debates were initiated by the minister of foreign affairs during February each year and presented the statement by addressing the issues. Following the discussion, the concerned representatives of the Advisory Council on Foreign affairs expressed their views and raised the issues on the statement. Since Sweden is a member of the European Union(EU), foreign policy deals with the coordination of the EU (RIKSDAG, 2021).

The importance of foreign policy is to safeguard the country and the people by making security policies based on cooperation and coordinating with other countries and other organizations. The key issues being discussed are improving Social Security, working in cooperation with neighboring countries to develop military collaborations, eradicating poverty, sustainable development, climate change to reduce emissions to avoid global security threats, feminist foreign policy to strengthen women's rights, maintaining gender equality, terrorism, migration, pandemic, racism, religion and human trafficking (RIKSDAG, 2021).

2.1.2 Budget Debate

Every year during September or October, the minister of finance submits the budget bill to parliament, followed by a discussion with the members of parliamentary parties. A further debate followed by the amendment of the Spring budget will proceed during April or June and amendment of Autumn budget rarely in November. The data contains eight debates for two with each financial year from 2014 to 2018. (RIKSDAG, 2021).

2.1.3 Interpellation debate

Parliament controls government strategies through interpellation. An interpellation is a question raised by the parliament member through writing or orally on a concerning issue to a government minister. Furthermore, concerned ministers are responsible to

reply within fourteen days both written and verbally. Through interpellation debate, parliament questions the Government's strategies (Magnusson, et al., 2018). All the members of parliament are welcome to participate in the debate. 249 interpellation debates that took place from 2014 to 2018 were further categorized according to the category of the ministry shown in table A (see Appendix A for Sweden Government ministries classification according to the ministers serving period during 2014-2018).

The table is formulated with the properties name of the government ministry, along with the responsibilities of each ministry with concerned minister name, party affiliation, and serving period. The members of the Sweden Government consist of a Prime Minister and Twenty-one ministers (Government, 2021).

2.1.4 Debates on Parliamentary Business

Parliament is responsible for taking decisions on various business issues recurrently every year, from budget bills to EU affairs. Parliamentary business debates are being discussed to take responsible decisions generally originated through the proposals from the various ministries of the government or the members of the parliament to check the performance of their duties (RIKSDAG, 2021). From 2014 to 2018, 201 debates are being discussed and categorized further according to the type of government ministry.

2.1.5 Current Debate

A special debate or current debate is the request from any party on their selected subject to Parliament. Then the speaker discusses with the party leaders and decides whether and when the debate can be arranged. The government minister is responsible for the current topic during the debate, and the party leader needs to participate (RIKSDAG, 2021). During the tenure of fiscal years 2014 to 2018, twenty-nine current debates have been arranged and categorized according to a type of ministry.

2.1.6 Corpus categorization

After the interpellation, parliamentary business and current debates were categorized according to the responsible ministry of the debate; the corpus contained 16 classes of 1632 documents. The number of documents that belong to every single activity is mentioned in table 2.

Table 2 : Categories of the Corpus

Documents Classes	Number of Documents
Climate and Environment	104
Constitutional Affairs	48
Culture and Democracy	74
Defense	57
Education and Research	123
Employment and Establishment	91
Energy and Digital Development	46
Enterprise and Innovation	67
Finance	218
Foreign Affairs	147
Gender Equality, Children and Elderly	42
Health and Social	147
Housing and Urban Development	84
Infrastructure	117
Justice	196
Rural Affairs	71

2.2 Properties of the Debates(Metadata)

The attributes of the debates needed for the analysis are extracted from the XML file and explained in table 3. The attributes are the metadata of the debates, and description is the detailed explanation of the fields.

Table 3 : Properties of Debate Document

Attributes	Description
Dok_id	A Unique ID, composed of period and category of the debate.
Dok_rm	Parliamentary period.
Kammaraktivitet	Chamber activity, labels the type of debate.
Avsnittsrubrik	Topic title, explains the specific topic of the debate.
Talare	Speaker name, labels the name, type of ministry and Party affiliation of the current speaker.
Parti	Speaker party, labels the party affiliation of the current speaker.
POS	Parts-of-Speech, labels each word with appropriate POS tag.
Lemma	Base form of the specific word.

2.3 Pre-processing of data

The tokenization step became simple with the data being enriched with semantics and split into words. The attribute lemma of the corresponding speech text is being used for the analysis, which helped to skip the stemming of the words.

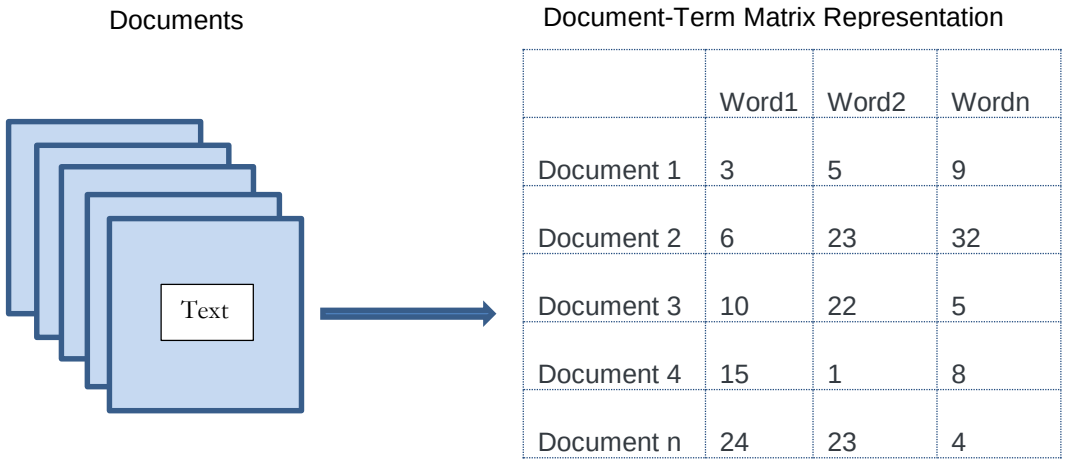
1. The lemma of each token available from the enhanced dataset is considered for the analysis.

2. The numbers, symbols, punctuations, and fewer than three character words were removed.
3. Remove the Swedish stop words which do not have substantial meaning using the Natural Language Toolkit(NLTK) stop words package.
4. Common prepositions which do not add meaning were filtered by utilizing POS tags of the enriched corpus.

As a result of the above steps, the corpus contains 5,804,760 tokens with 1632 documents.

2.4 Transformation of Tokens

Topic modeling algorithms take data in the form of a matrix. The matrix is formed by transforming the pre-processed tokens into a frequency matrix by counting the document's words. The matrix is termed DTM. DTM is a BoW representation that does not consider the order of the words within a document. Figure 1 visually explains the transformation of documents into DTM how text is turned into a matrix representation. The rows correspond to the corpus documents, and columns correspond to the terms of the documents. The words of each document are counted and stored in the corresponding cells.



	Word1	Word2	Wordn
Document 1	3	5	9
Document 2	6	23	32
Document 3	10	22	5
Document 4	15	1	8
Document n	24	23	4

Figure 1 : DTM Representation

The parliamentary debates contain many repetitions of words by the speakers. At first, as a standard setting, a DTM is created by using the `cast_dtm()` function from the `tidytext` package. The term frequency-inverse document frequency(TF-IDF) weighting factor (Blei & Lafferty, 2009) is applied to the DTM to trim the number of terms. This weighting factor generates word statistics based on its frequency in a document using the `bind_tf_idf()` function from the `tidytext` package. The terms that are lower than the median of the TF-IDF factor are removed, ensuring the less essential words across the documents. A matrix of 1632 documents was constructed into 23064 terms during the transformation step.

2.5 Implementation of LDA

In topic modeling, the implementation of LDA successfully processes extensive data due to its nature of exchangeability property (Blei, et al., 2003). The algorithm treats each document as a random mixture of latent topics, characterized using the multinomial distribution of words (Blei, et al., 2003). The distribution of topics among documents is labeled as Dirichlet distribution (Blei, 2012). This process makes the model understand the given corpus as a heterogeneous text and summarize the outcome by overlapping all corpus terms (Blei & Lafferty, 2009). As a result, LDA generates a weighted list of topics for the corpus. (Campbell, et al., 2015).

LDA was performed on the tokens of the documents that had gone through the preprocessing and transformation phases. There are several implementations of the algorithm. Here two of them are chosen to explore in-depth.

At first, the model was implemented with topic models and the tm package for interpreting the results with posterior probabilities of word and topic distributions. Furthermore, the algorithm was developed with `FitLdaModel()` function from the `textminer` package. That package provides the metrics for better model evaluation by giving log-likelihood, coherence and R2 metrics.

The steps for processing the algorithm are depicted in figure 2, where M documents with N-words were fed into the model. According to the fitted model, the output was several topics(k) as a cluster of words represented by phi(**ϕ**) and the topic proportions of every single document represented by theta(**Θ**). At first, the fitted model initializes the concentration parameters alpha(α) and beta (β) parameters and randomizes the topic proportions using the Bayesian approach. A Gibbs Sampling was then utilized to resample words per topic(Z) and produce results(k) after the iterations were completed.

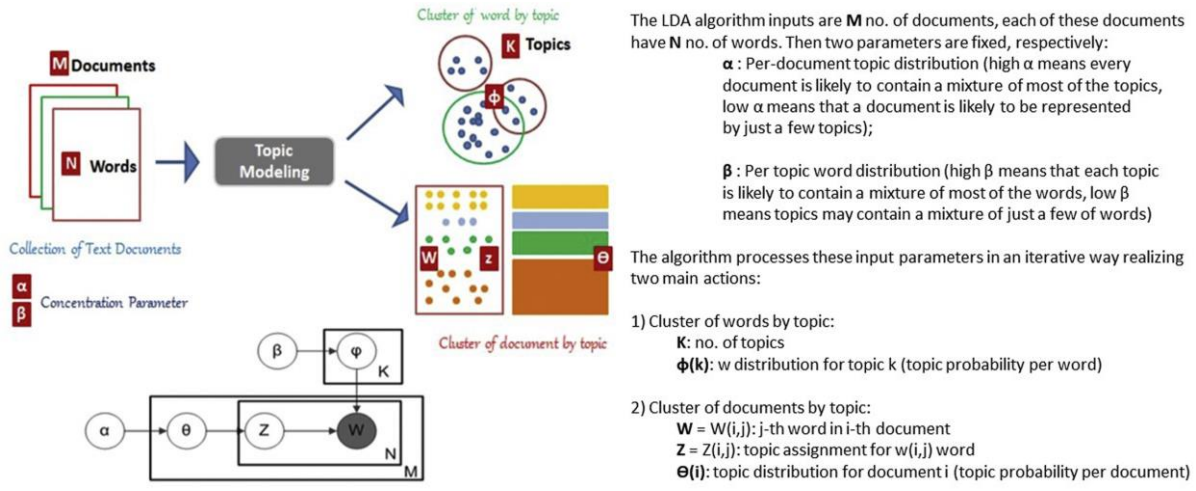


Figure 2: Graphical representation of LDA *(Rortais, et al., 2021)*

2.5.1 Arguments in LDA Model

Document-term matrix(DTM): The first parameter is the DTM formed during the transformation phase of the corpus.

Number of Topics(k): The algorithm takes k as the input, the number of topics needed to extract from the corpus. The current implementation has the advantage of knowing the number of topics and the type of ministry to which the document belongs.

Cal_loglikelihood: The metric evaluates the goodness of the fitted model by taking the log of tokens per topic probabilities. For every 10 iterations, log-likelihood is stored into the fitted model (Jones, et al., 2016).

Iterations: The algorithm takes the number of iterations as input to guide the model about the repetition of the Gibbs sampling (Blei, et al., 2003). By iteratively running the model with distinct iterations 200, 500, and 1000 the number of iterations is fixed as 300 by comparing the Log-likelihood. The optimal number of iterations is when the likelihood is greatest and not convergent with further iterations (Blei, et al., 2003). By plotting the number of iterations on the x-axis and corresponding log-likelihood on the y-axis, as shown in figure 3, observed that at 300 iterations, the log-likelihood reached approximately maximum value.

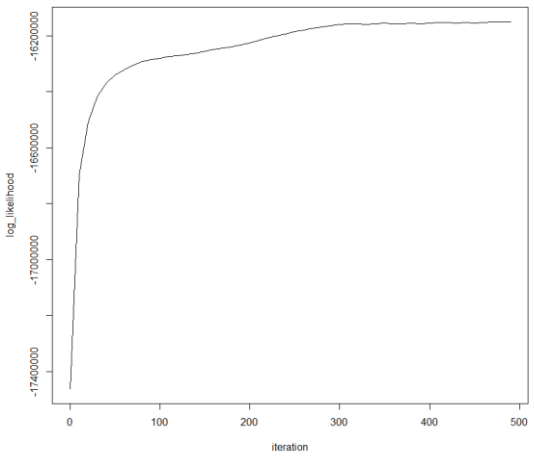


Figure 3: Exploration of number of iterations based on log likelihood

Alpha(α): Prior distribution factor controls the topics over documents that evaluate the document's density. The factor initializes by default to 0.1 (Blei, et al., 2003). High α value explains the document as the mixture of most of the topics, and a low value indicates few topics.

Optimize_alpha: During the implementation, chosen to optimize alpha for every 10 Gibbs sampling by setting the parameter as true.

Beta (β): The prior distribution factor controls the words over topics. The factor evaluates the topics word density, termed as a per-topic-per-word probability, and controls the distribution of words per topic in the document. A high β value explains that the topic contains a mixture of most of the words and a low value indicates less number of words.

Cal_r2: Calculates the R-squared value on the fitted model that explains the degree of variability of the topics, by setting the parameter as True.

Cal_Coherence: Coherence is the qualitative probabilistic measure that calculates by taking a vector of topics with top five terms by default and DTM. This factor explains the quality of the topics by considering the amount of semantic resemblance between the top five words within the topic (Jones, et al., 2016).

2.5.2 Resulting factors of fitted Model

Beta (β): The algorithm's outcome holds the beta parameter as the posterior probabilities of terms for the topic models package. The slice_max() function from the dplyr package generates top words from each topic. These words are then analysed, interpreted and depicted in figure 12.

Gamma(γ): Documents topic density factor termed as a per-document-per-topic probability. This factor contains the estimated proportion of words from the documents generated from that topic. Topics most associated with documents are interpreted and shown in the results (see figure 13), using the slice_max() function of the dplyr package. Documents topic density factor termed as per-document-per-topic probability.

Phi (ϕ): This parameter explained the word proportions of each topic based on topic probability per word.

Theta(Θ): This parameter explained the topic proportions of each document based on topic probability per document.

Prevalence: The value measures the probabilistic topic distributions to explain the frequency of topics within the corpus of documents (Lee & Mimno, 2014). This distribution was calculated by taking the column sum of theta for the fitted model, dividing by the total sum of theta, and finally multiplying by a hundred.

At first, the LDA model is built with a document-term matrix, number of topics, number of iterations and seed number to replicate the result with the topicmodels package. The model's probabilistic beta and gamma outcome factors are interpreted using functions from the tidytext package. At first, the number of topics is 5 with five debates original unclassified data. Further changed the number of topics to 16 and applied LDA on classified data.

Furthermore, using the textminer package, LDA was implemented on classified data with arguments DTM, the number of iterations, optimize_alpha set to 'true', cal_coherence set to 'true', cal_loglikelihood set to 'true', cal_r2 set to 'true'. The resulting topics and the quality metrics of the fitted models were discussed in chapter 3.

2.6 Implementation of STM

The STM improves the models by enabling variables of interest to be included in the prior distributions for document-topic proportions and topic-word distributions. Rather than assuming that the topical prevalence (the frequency with which a subject is discussed) and topical content (the words used to discuss a topic) were constant across all participants, the analyst might include factors that we might expect to explain the variation in the data. As a result, each open-ended response in the model was a combination of subjects.

Topic proportions (θ) can be associated with the STM. The prevalence of those topics can be changed by a collection of variables X through a typical regression model with covariates $\theta \sim LogisticNormal(X\gamma, \Sigma)$. A topic ($\mathcal{Z}$) is chosen from the response specific distribution for each word (ω) in the response, and a word is chosen conditionally on that topic from a multinomial distribution over words parameterized by β, which is formed by deviations from the baseline word frequencies (m) in log space ($\beta\kappa \propto \exp(m + k\kappa)$).

Besides, another key responsibility of STM was that rather than sharing a global mean, each document had its prior distribution over subjects, specified by covariate X, and the use of words within a subject can be affected by covariate ∪. These extra factors allow the topic model's prior distributions to be "structured," adding helpful information to the inference method.

The STM allows for quick, clear, and repeatable analyses with little a priori assumptions about the texts being studied. Although it is a computer-assisted approach, the researcher remains an essential element in comprehending the texts, as demonstrated in the Examples section. The model and the texts themselves govern the analyst's interpretation attempts. However, as we demonstrate, the STM may relieve the analyst of the burden of having to create a categorization system from the start (Lee & Mimno, 2014) and execute the frequently tedious task of associating documents with those categories.

A graphical depiction of the model is shown in figure 4. The generative model starts at the top, with document-topic and topic-word distributions producing documents with metadata (X_d, where d indexes the documents). A topic is defined as a combination of words with a probability of belonging to a subject in this framework (which is the same as other topic models like LDA without the metadata). A document is a collection

of themes, which means that a single document can include several subjects. As a result, the total of all subject proportions for a text is one, as is the sum of word probabilities for a particular topic.

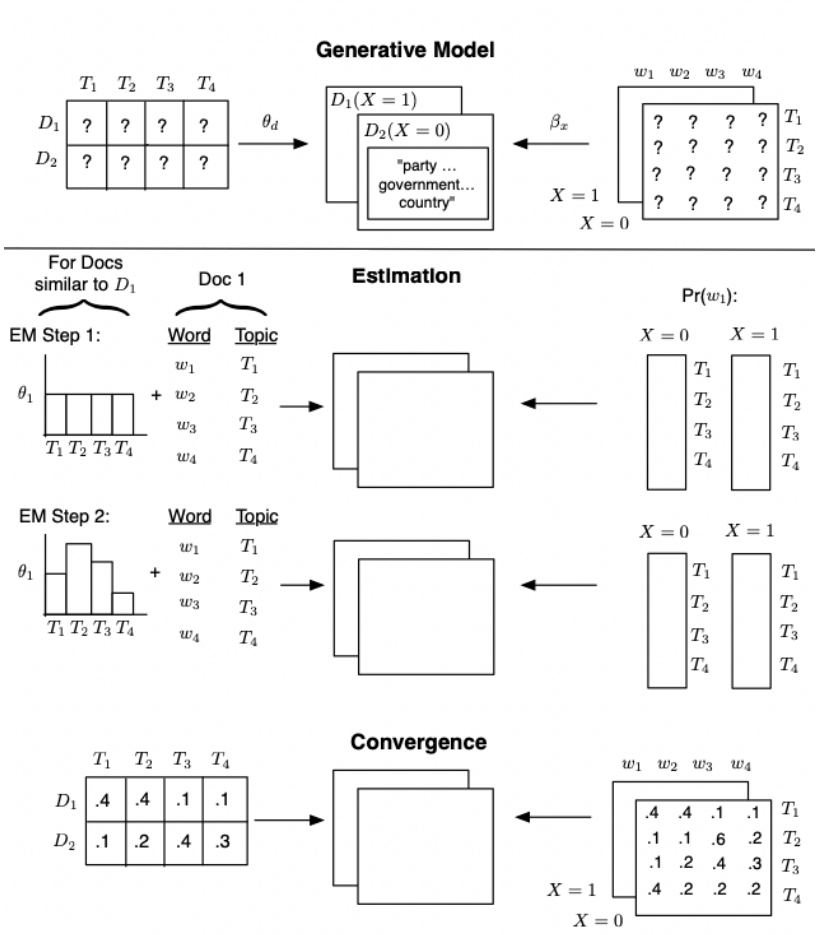


Figure 4: STM estimate and heuristic description of the generating process.

As shown in figure 4, the document generating process and the statement below illustrates the scenario where topical prominence and content can function as document metadata. Topical prevalence refers to the percentage of a document related to a subject (explained on the left), while topical content refers to the words used inside a topic (described on the right-hand side).

Topical prevalence covariates are hence metadata that explain topical prevalence, whereas topical content covariates are variables that describe topical content. However, the model allows topical prevalence variables, topical content covariates, both, and neither. The model simplifies to a (rapid) implementation of the Correlated Topic Model when no covariates are present (Lee & Mimno, 2014).

2.6.1 Arguments in STM Model

Documents: The term matrix for the document to be modeled. The native STM format, a sparse term count matrix with one row per document and one column per term, or a quanteda DFM (document-feature matrix) object can all be used. This will include the vocabulary and, in the case of quanteda, potentially the metadata when utilizing the sparse matrix or quanteda formats. After cleaning the data (Discussed in the Transformation of Token) we converted the DTM to DFM using quanteda. Moreover, the prepDocuments API from the tm package was used to re-index all metadata/document relationships to input into the STM document.

Vocab: The words in the corpus are listed in the order of their vocab indices in documents in this character vector. Each phrase in the vocabulary index must occur in the documents at least once. We used vocab from the documents after converting DTM to DFM with “STM” classifier.

K: The intended number of subjects is usually represented by a positive integer (of size 2 or more). The current implementation has the benefit of knowing how many topics were covered before and the type of ministry to which the document belonged.

Prevalence: This is a matrix comprising topic prevalence covariates or a formula object with no response variable. As discussed in the methodology, we used party affiliation and speech giver’s position as covariant from the document metadata.

Data: a data frame with the prevalence and/or content covariates as an option. The variables are taken from the active environment if they are not given. We gave our metadata as data from the document frequency matrix.

Verbose: Whether information should be printed on the screen is indicated by this logical flag. Verbose was used to verify the process in the console.

2.6.2 Resulting factors of fitted Model

Beta: For each topic, a list contains the log of the word probabilities. The top words generated from each topic, using the slice_max() function from the dplyr package, are analyzed in figure 17.

Convergence: The value of the approximate bound on the marginal probability at each stage is included in the list of convergence elements. We used convergence bound for calculating our log-likelihood.

Theta: Matrix showing topic proportions based on the number of documents vs the number of topics. The proportions of topics based on the number of documents vs the number of topics, using the slice_max() function from the dplyr package, is shown in figure 18.

2.7 Sampling distributions and the Bootstrap

A non-parametric bootstrap framework was implemented to assess the uncertainties in sample distributions by calculating the mean confidence interval statistics. The bootstrap sampling technique draws inferences from original data with random samples (Zivot, 2017). Through simulation, the process treats samples of each iteration as the original population to inherit the distribution of the original dataset (Danielsson, et al., 2001).

At first, bootstrap with boot package written by Davison & Hinkley (1997) was implemented where bias and standard deviation were considered for calculating the estimated statistics. Due to the sampling observations needed to draw using indices, the dataset structure is unsuitable for utilizing the in-built package's benefits. Furthermore, the non-parametric bootstrap resampling process was implemented through a custom function, where all parts of the program were included inside a loop and iterated through several samples. The implementation steps are as follows:

1. **Resampling Data:** Generate N random samples with replacement from the original data. In this way, each generated sample has the same number of observations as the original data.
2. **Estimate parameter:** With each sample, a parameter was estimated; as a result, the N number of observations for the parameter was saved.
3. **Compute Confidence interval:** A 95% confidence interval(CI) is computed to test the hypothesis on the saved parameter. The CI statistic explains the degree of accuracy on the estimated parameter (DiCiccio & Efron, 1996).

The LDA and STM models have been looped through 200 iterations at once through custom function to get reliably estimated parameters. Grouping activities with maximum topic proportions summarized 3200 observations to calculate the mean gamma proportions of 16 activities of parliamentary ministries. Further, uncertainties were assessed by constructing mean confidence intervals on the parameter, mean gamma probability. Finally, the null hypothesis was explained by performing a t-test on the original data and sampling distributions by observing the significant p-value.

Chapter 3

Results

3.1 Visualizing data with Word Cloud

We began the analysis using `wordcloud` (Fellows, et al., 2018), to observe the most frequently used words during the debates. A wordcloud displays text size in proportion to its frequency within documents (Kwartler, 2017). A preprocessed DTM was created for five individual debates then respective word clouds were plotted, shown in figures 5,6,7,8 & 9. Important words concerning the corpus were examined visually to understand the corpus better.

Figure 5: WordCloud for Budget Debate

Figure 6: WordCloud for Foreign Policy Debate

Figure 7: WordCloud for Special Debate

Figure 8: WordCloud for Debates Parliamentary Business

Figure 9: WordCloud for Interpellation Debate

The word cloud for foreign policy showed relevant words of international organizations, security, religion, improving the job market, and climate coordinating the neighboring countries. Furthermore, the word cloud for finance showed the most representative words: the job market, improving economy, and tax issues. Both of these debates depicted that documents were related to their concerned ministry. On the other hand, the Special Debate, Debates on Parliamentary Business and Interpellation Debate, discussed the most diverse subjects, which was the primary reason to further classify the corpus according to the debate headline mapping to the ministry of activity.

3.2 Descriptive statistics of the Corpus

The classified corpus from 2014 to 2018 was visualized using the geom_line() function. A line plot in figure 10 represents the number of documents per financial period. The x-axis indicates the debates financial period, while the y-axis indicates several documents. Each line represents one activity of the ministries labeled at the end of the line.

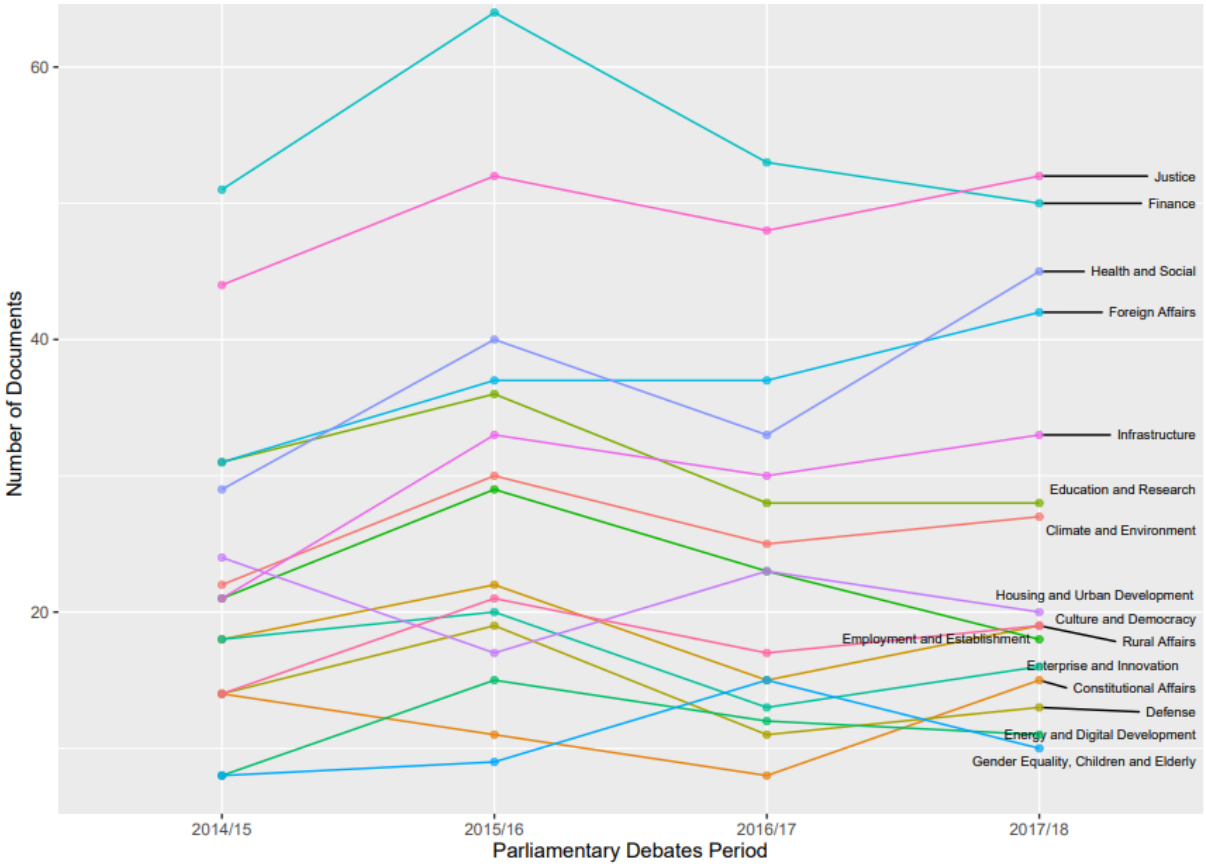


Figure 10: An overview of the number of documents of the corpus according to the concerned ministries per parliamentary period

According to the plot, the documents for each ministry were available during four parliamentary periods. Though the ministries of Energy & digital development and Gender Equality, Children & Elderly provided a smaller number of documents, 46 and 42 respectively, compared to Finance and Justice yielded more 281 and 196 documents, the sample size was reasonable for topic modeling (LDA, and STM) and no numerical issue was encountered in model fitting.

3.3 Results of LDA

At first, the topics of an unclassified five debates (table 1) were fitted, and the model's performance was depicted using a boxplot (see figure 11). The resulting topics of the fitted model are mapped to their related activities using gamma probability measure. The model generated five topics are on the x-axis, and the percentage of terms of these topics that belong to a particular type of (actual) parliamentary debate is shown on the y-axis. From the distribution of topics, it is clear that budget and foreign policy debates are correctly identified with a median of 70% and 80%, respectively. The other three topics - Aktuell Debate, Interpellations Debate and Ärende Debates all overlapped among all five topics. This overlap led to further data classification by relating the subject of debate with the responsible government ministry. Thus the total corpus of documents was classified into 16 classes (table 2), and therefore used the number of topics as 16 in the second implementation of LDA.

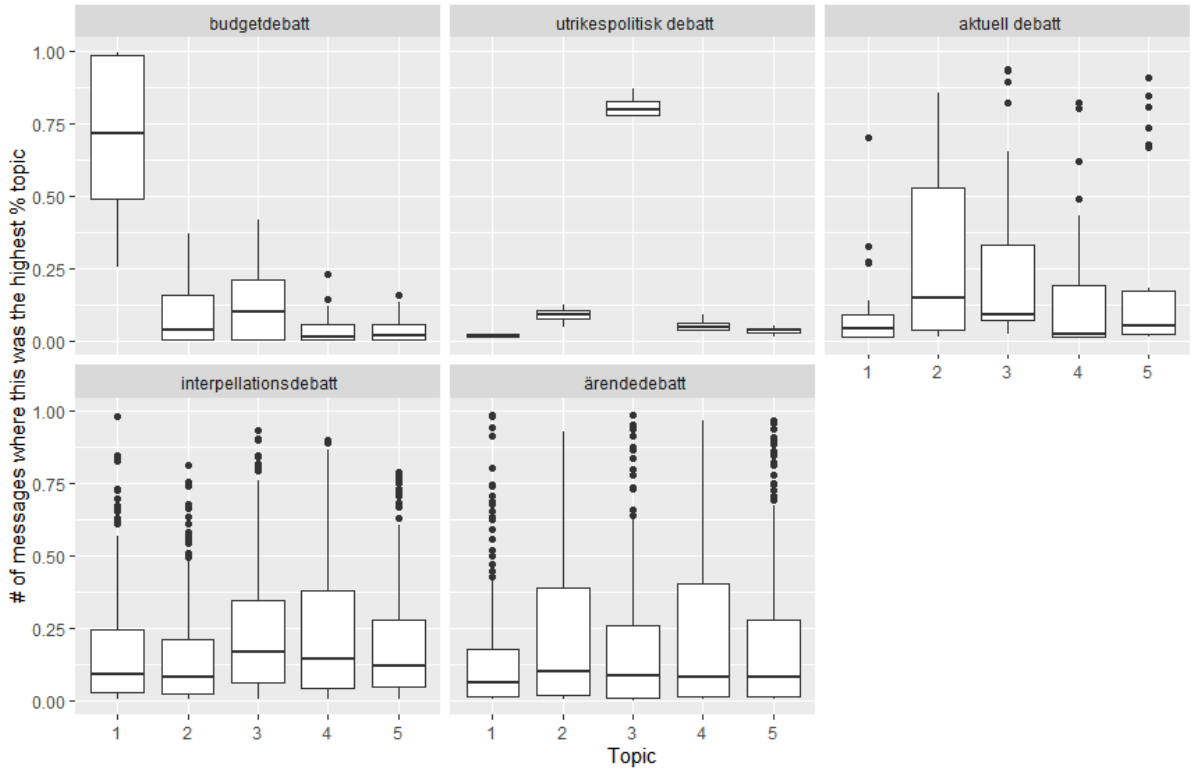


Figure 11: Distribution of topics over debates.

Then, 16 topics and their associated terms of the fitted model on classified data were visualized. The resulting terms were extracted using beta posterior distributions and visualized using the `ggplot` package shown in figure 12. The top ten terms of 16 topics are displayed on the y-axis according to the beta posterior probabilities shown on the x-axis. By observing manually, the words related to most of the topics are distributed into a unique subject.

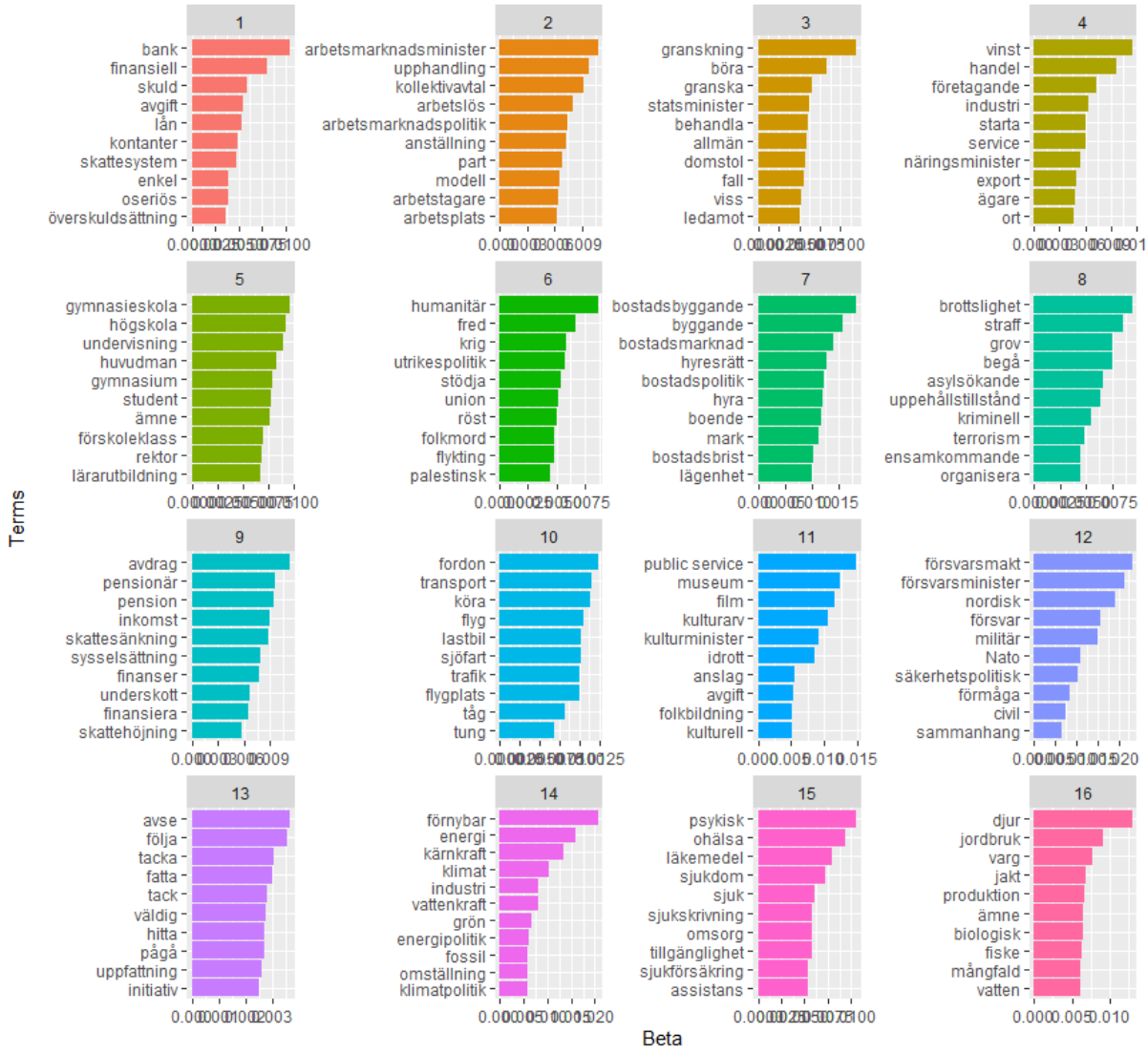


Figure 12: Top ten terms of 16 Topics (obtained from LDA model).

Furthermore, the topic proportions belonging to the respective activity documents were extracted using gamma posterior distributions and shown using a boxplot in figure 13.

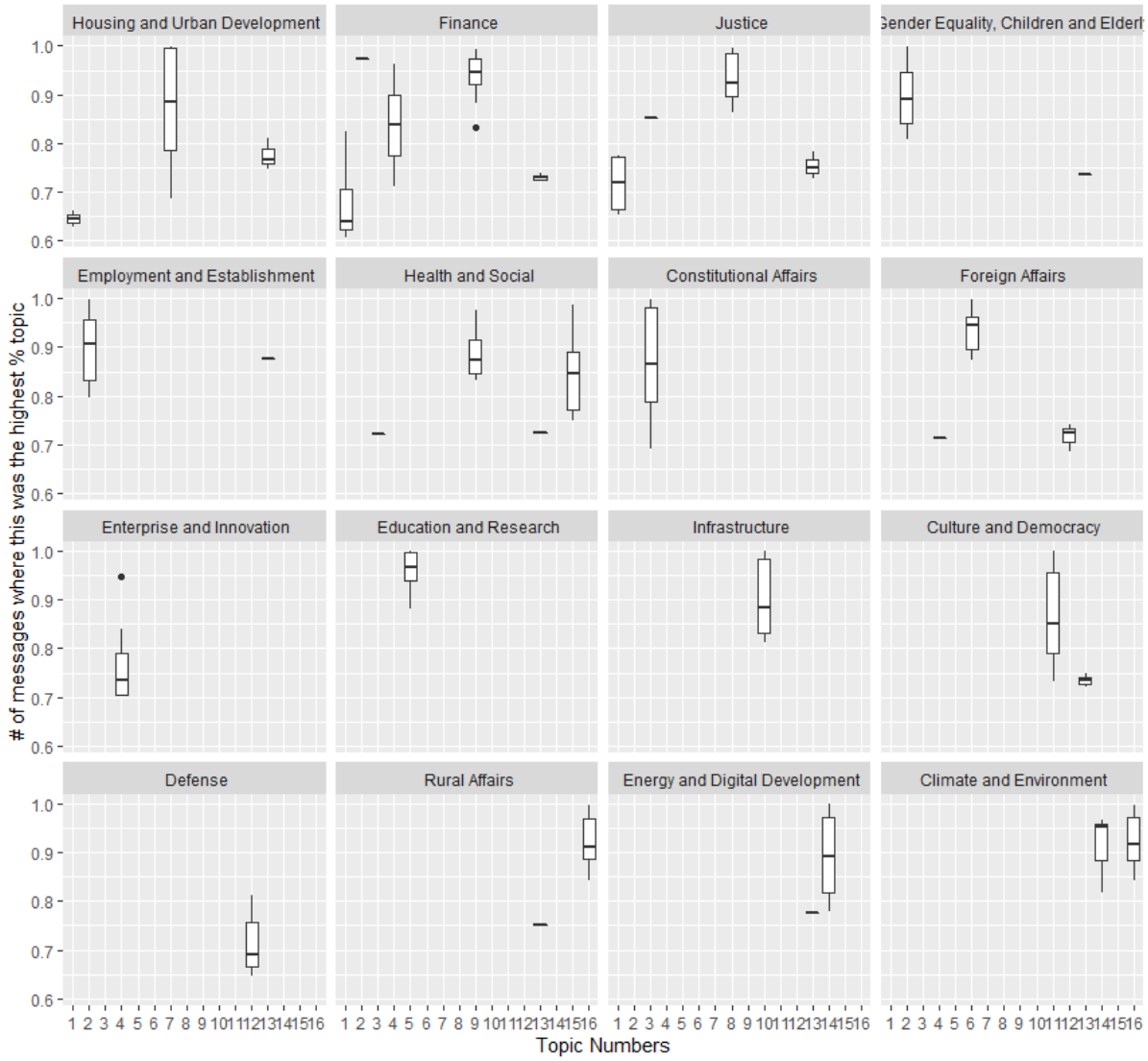


Figure 13: Distribution of topics over classified debates (top terms).

By comparing the words per topic from ggplot and its concerned boxplot following observations were depicted. Topic 1 is a mixture of Housing & urban development, Finance, and parts of Justice ministries. The topic can be explained by financial issues occurring under Housing and Justice ministries. Topic 2 is explaining Employment & Establishment, Finance and Gender Equality ministries. This explains that, during the financial debates, the discussion regarding the improvement of labor is critical. Gender Equality is discussed mainly in the labor market. Topic 3 is constitutional affairs explaining the critical parliamentary decisions, a part of the Justice ministry. Topic 4 is a mixture of Enterprise & Innovation, Finance, and foreign affairs ministries. This explains that during the financial and foreign debates, industrial issues are one of the critical factors. Topic 5 is independently classified as the Education and Research ministry. Topic 6 is classified as the Foreign Affairs ministry. Topic 7 is classified as the Housing and Urban Development ministry. Topic 8 is classified as the Justice ministry. Topic 9 is a mixture of Finance and Health & Social ministries. Topic 10 is classified as the infrastructure ministry. This topic is mainly classified as financial issues of pensioners. Topic 11 is classified as the Culture and Democracy ministry.

Topic 12 is a mixture of foreign affairs and Defense ministries. This topic is mainly classified as a military issue. Topic 13 words explain the adjectives which are not interpreted as belonging to any ministry. Topic 14 is a mixture of Energy & Digital development and Climate & Environment ministries. Here the topic is explained by factors of energy consumption and climate saving. Topic 15 is classified as Health and Social ministry. Topic 16 is a mixture of Rural affairs and Climate & Environment ministries. The topic is mainly explaining the environmental concerns related to rural issues. The coherence and prevalence scores of the topics are listed in table c1 (see appendix C).

Then using the LDAvis package, the outcome of the LDA model was visualized in figure 14. The package supports an interactive web-based visualization of global topics and their respective terms, along with their frequencies. This interactive technique took input from fitted LDA models by considering posterior probabilities of topics and terms. A JSON object was created to serve the interactive visualization. This visualization technique mainly has two panels. First, the left panel shows a global view of the topics. The topics are drawn as circles in the two-dimensional plane in this view. Here the center was determined by computing the distance between topics (Mei, et al., 2007). The circles were placed by considering the prevalence of the respective topics. Second, the right panel shows a horizontal bar chart of their terms for the selected topic on the left (Mei, et al., 2007).

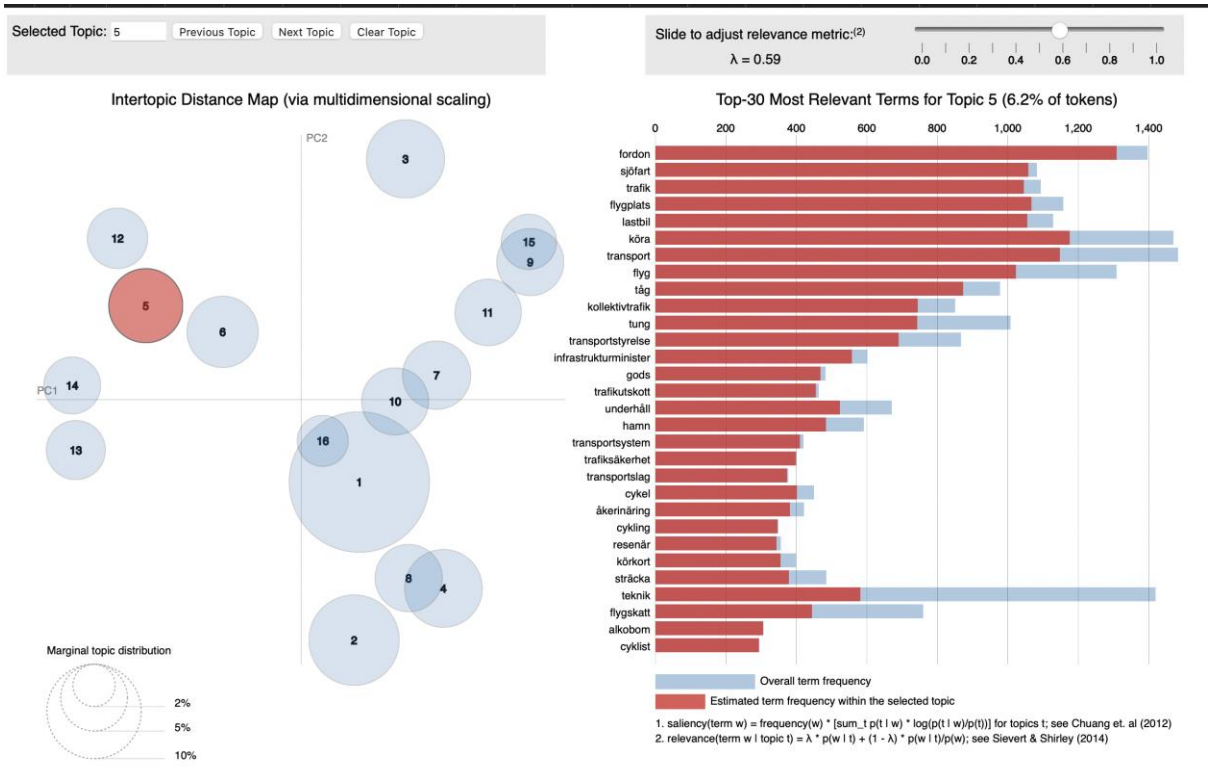


Figure 14: The global view of LDA topics on the left, and their respective term bar charts are on their right (with topic 5 is selected).

Using the LDA fitted model, topic 5 was selected from the left panel, and the respective terms with the bar chart were shown on the right panel. The topic numbers got rearranged since LDAvis considered posterior probabilities of terms and topics, and the previous boxplot took gamma probabilities. The most probable terms could interactively adjust with the relevance (ʎ) measure. The optimal value of ʎ 0.6 shows accurate terms (Mei, et al., 2007). The left panel shows an overlap between the topics, especially 1 & 16, 9 & 15, 4 & 8, and 7 &10.

3.4 Results of STM

3.4.1 Model comparisons and determination of K

We start with the most important choice we will make when generating topic models: deciding on a reasonable number of topics K. We focused on the target of inquiry or the text corpus. For small corpora with highly narrow objectives of the research, the STM package authors propose 3-10 topics, 5-50 topics for corpora with a few hundred to a few thousand documents.

We start by making a table that contains all the various K values we intend to employ. In our scenario, we compute four models with 6, 16, 26 and 36 themes.

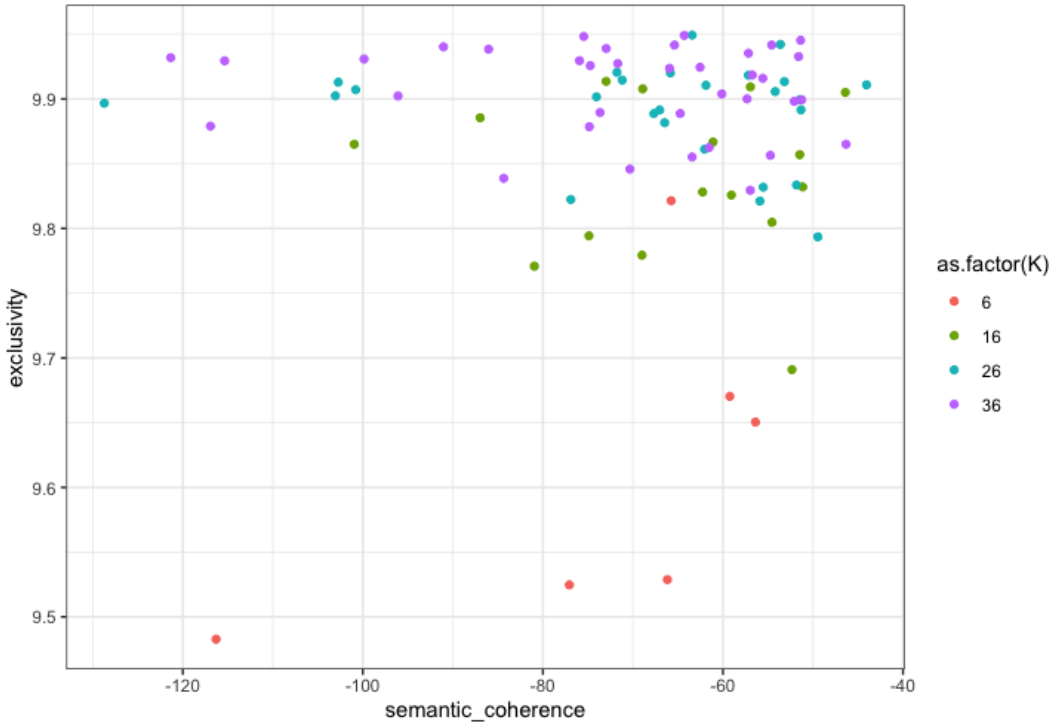


Figure 15: The results of selectModel K showing each model's average by numbers, while dots reflect topic-specific ratings.

At first, from figure 15, the models appear to be similar; nevertheless, we can observe a prominent outlier topic with much worse semantic coherence in the model with K = 6, and there appear to be some concerns with comparatively less exclusivity in all

outlier topic with much worse semantic coherence in the model with K = 6, and there appear to be some concerns with comparatively less exclusivity in all models. We determine the mean value of the various characteristic values to compare the models more quickly.

The typical trade-off between semantic coherence and exclusivity can be seen quite clearly here: the model with the lowest exclusivity has the lowest semantic coherence (K = 6); on the other hand, both K = 16 have the high exclusivity and the lowest semantic coherence, but they are very similar on both dimensions. As a compromise, we offer K = 16 as the model we will now use, classified manually.

3.4.2 Exploring words associated with topics

According to Mei, et al., (2007), Topic labeling in general aims to assign one or a few descriptive phrases to represent each topic and generate such phrases automatically.(Mei, et al., 2007)) went further to mention that the methods usually involve two main steps: (1) the generation of candidate labels (e.g. text or images) for a given topic; and (2) the ranking of candidate labels by relevance to the topic. Textual labels have been sourced in several ways, including noun chunks from a reference corpus. The labelTopics() function may investigate and explore the words connected with each subject. SageLabels may also be utilized in models that incorporate a content covariate. These functions will output to the console terms linked with each topic.

The function outputs a variety of word profiles by default, including highest probability words and FREX words. Words are weighted in FREX based on their overall frequency and uniqueness to the topic. Lift assigns more significant weight to terms that appear less often in other subjects by dividing their frequency into other topics by their frequency in other topics.

labelTopics (model_stm, c (3,9)) Topic 3 Top Words: **Highest Prob:** järnväg, bil, fordon, köra, transport, flyg, lastbil **FREX:** trafiksäkerhet, alkobom, cyklist, farledsavgift, infrastrukturproposition, cykelstrategi, trafikant **Lift:** godsstrategi, pråmtrafik, godståg, höghastighetsjärnväg, infrastrukturproposition, trafikant, trafiksäkerhet **Score:** sjöfart, järnväg, infrastrukturminister, trafiksäkerhet , transportstyrelse, alkobom, cyclist
Topic 9 Top Words: **Highest Prob:** patient, psykisk, vinst, assitans, läkemedel, ohälsa, onsorg **FREX:** patientsäkerhet, sjukvårdslag, ungdomspsykiatri, patient, patientlag, assistans, vårdgivare **Lift:** vårdkedja, akutbesök, äldrelag, alkoholförsäljning, alkoholfråga, alkoholpolitisk, almänläkare **Score:** patient, assitans, primärvård, apotek, läkemedel, psykisk, patientsäkerhet

Figure 16: Output of words associated with topics

The stmpackage's labelTopics() method provides the first summary, indicating essential terms based on four criteria for each topic. For clarity, only the first two topics are presented in the course view; if we execute the lableTopic(model_stm), we should receive a lengthy output with all 16 subjects. Looking at Greatest Prob, according to (T.Hofmann, 1999), refers to the word probability per subject (also referred to as β), and in this research, each topic receives seven words with the highest probability. The other three metrics are alternative calculations of the most meaningful words per topic; for example, FREX, as noted by (Hausser & Strimmer, 2009) (is for frequency exclusivity), tries to identify those words that are particularly distinctive for a given topic, as they frequently occur in the topic under consideration as well as in other topics (a word can namely, a high level on various topics as well β exhibit). Score and Lift also add more weights. The interpretation would be based on many or all four metrics in a perfect world. In terms of content, as shown in figure 16, we see topics that can be relatively clearly assigned to a thematic umbrella term - talks that contain topic 3 deal with infrastructure ministry and topic 9 with Health and Social ministry.

Initially, we fitted the STM model with party affiliation as a covariate; however, the resulting topics did not improve model fit. The resulting log-likelihood of the STM fitted model is -18186315.5, which is far lower than the LDA fitted model value of -16200000.

Then we found another crucial variable, the speech giver's position, which was constructed by mapping the speech giver's name with available roles in the parliament data. The resulting position property contained 31 classes (see Appendix B of table B), and additionally, we fit the STM model with covariate as speech giver's position. Using the fitted model, we obtain a log-likelihood of -14219018.95, significantly higher than the LDA of -16200000.

Similar to LDA, the resulting terms of the STM fitted model were extracted using beta probabilities and depicted in figure 17. The terms of each 16 topics were shown on the y-axis according to their beta factor on the x-axis using the `ggplot` package.

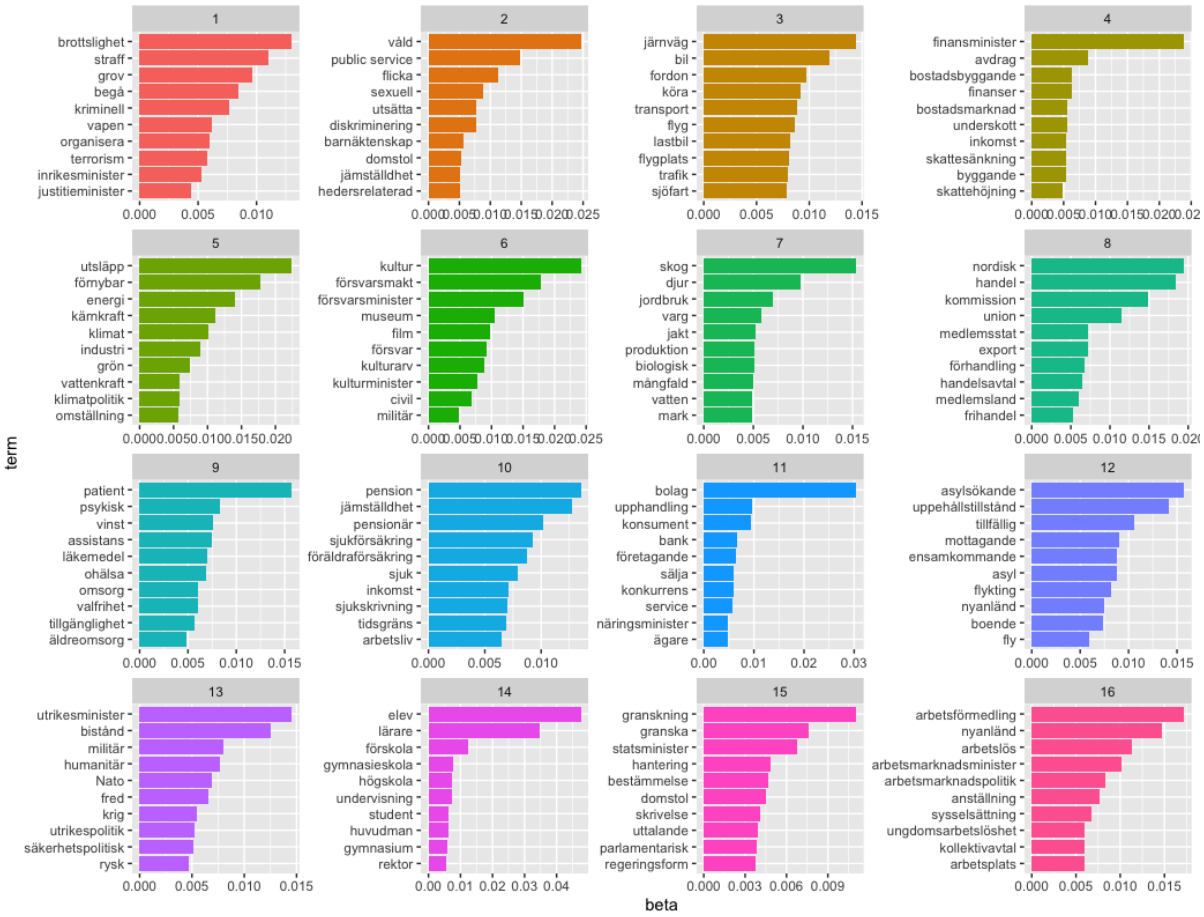


Figure 17: Top ten terms of 16 Topics (obtained from STM model).

Furthermore, the gamma distributions of documents are mapped to their corresponding activity and depicted using a boxplot in figure 18.

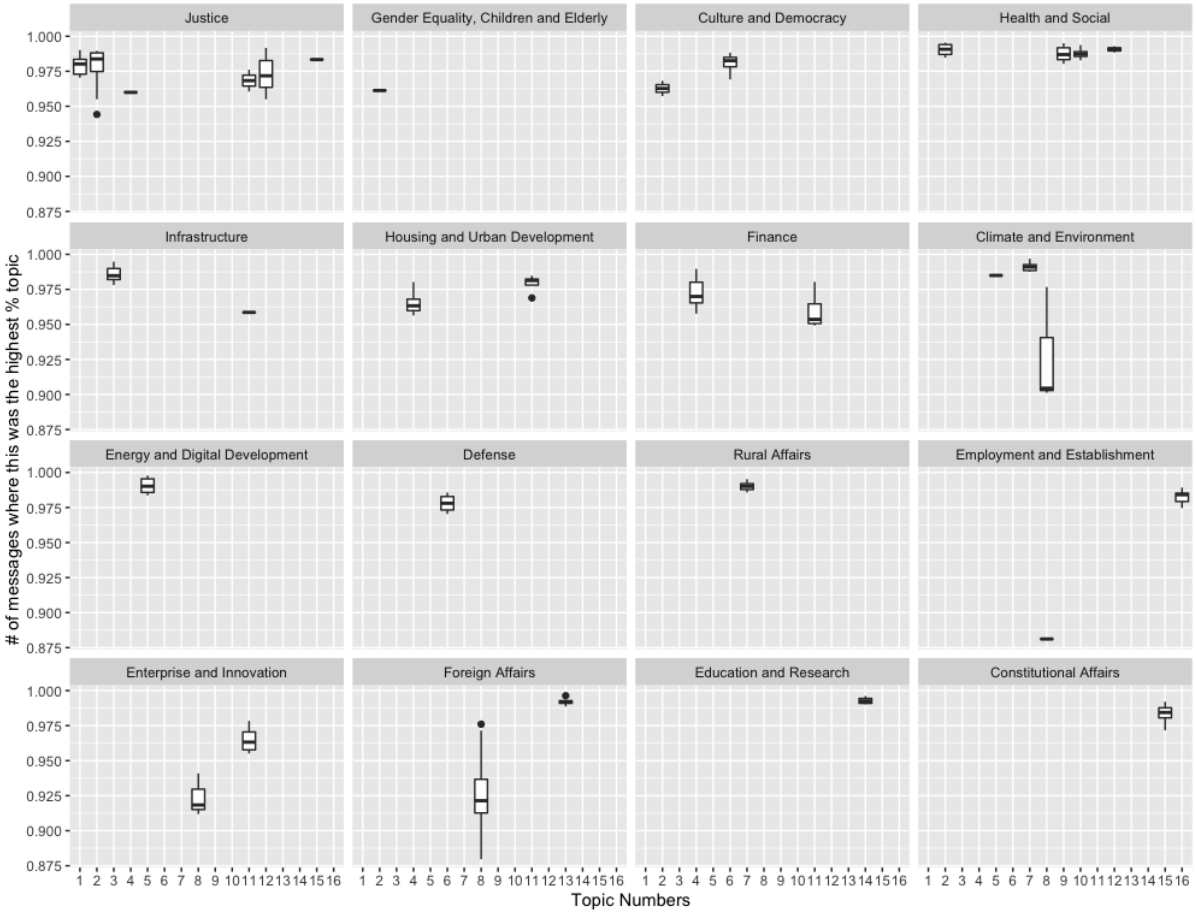


Figure 18: Distribution of topics over classified debates (Top terms)

The following observations were depicted by comparing the words in each topic from ggplot and its concerned boxplot with topic proportions. Topic 1 is classified as Justice. Topic 2 is a mixture of Gender Equality, Children & Elderly, Justice, Culture & Democracy and Heal & Social ministries. Topic 2 explains Employment & Establishment, Finance and Gender Equality ministries. Topic 3 is independently classified as the infrastructure ministry. Topic 4 is a mixture of Housing & Urban Development, Finance and Justice ministries. Topic 5 is a mixture of Climate & Environment and Energy & Digital Development ministries. Topic 6 is a mixture of Defense and Culture & Democracy ministries. Topic 7 is a mixture of Rural affairs and Climate & Environment ministries. Topic 8 is a mixture of Enterprise & Innovation, Foreign affairs, and Climate and Environment ministries. Topics 9 and 10 are classified as Health & Social ministry. Topic 11 is classified as the Justice, Infrastructure, Enterprise & Innovation, Housing & Urban Development, and Finance ministries. Topic 12 is a mixture of Justice and Health & Social ministries. Topic 13 is independently classified as foreign affairs. Topic 14 is classified as the Education and Research ministry. Topic 15 is a mixture of Justice and Constitutional affairs ministries. Topic 16 is classified as the one concerning the Employment and Establishment ministry. The coherence and prevalence scores of the topics are listed in table c2 (see appendix C).

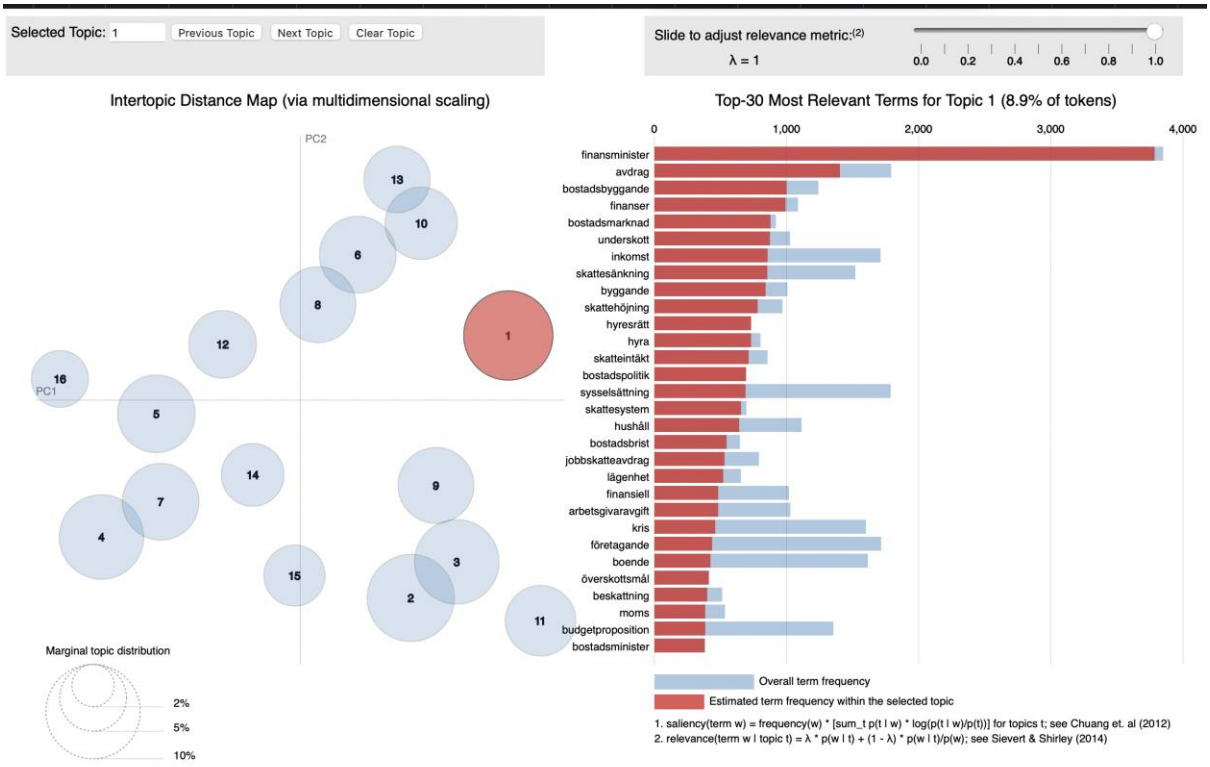


Figure 19: The Global view of STM topics on the left, and their respective term bar charts are on their right (With topic 1 is selected)

The toLDAvis plot was depicted in figure 19 using the fitted STM model. Topic 1 was selected from the left panel, and the respective terms were shown on the right panel with the bar chart. The topic numbers got rearranged since toLDAvis considered posterior probabilities of terms and topics, and the previous boxplot took gamma probabilities. The left panel shows that the maximum number of topics were separated individually compared to LDA, topic numbers 1, 6, 9, 11,12, 14,15, and 16. At the same time, there exists less overlap between the topics compared to LDA. The maximum overlap between the topics is only between two pairs, especially 2 & 3, 10 & 13. As a result, these observations showed that the STM model improved the topic estimation of 16 classes compared to LDA.

Then we ran an STM model after recoding the party affiliation categorized as far-left (Vanster(V), and Green(G)), far-right (Social Democrats(SD), and Kristian Democrats (KD)), and others (Moderates (M), Liberals(L), et al.) and provided categorized party affiliation as a covariate to see if this improved the fit of the model. Log-likelihood was calculated using the VEM approach (Hornik & Grün, 2011) on LDA and STM.

The resultant log-likelihood of LDA, STM with covariate speech giver's position and STM covariate with categorized party affiliation as a covariate were plotted in figure 20 to compare the performance of the models. As the procedure proceeds, it gets closer to the maxima very quickly. After 150 iterations, the likelihood has not changed

significantly. The likelihood evaluated at the optimum (e.g. posterior mode) is relevant when comparing two models. Here LDA showed a lower value of -16200000 when compared with both STM models. This explains the estimation effect of STM with the speech giver's position, and categorized party affiliation extracted more meaningful results than LDA.

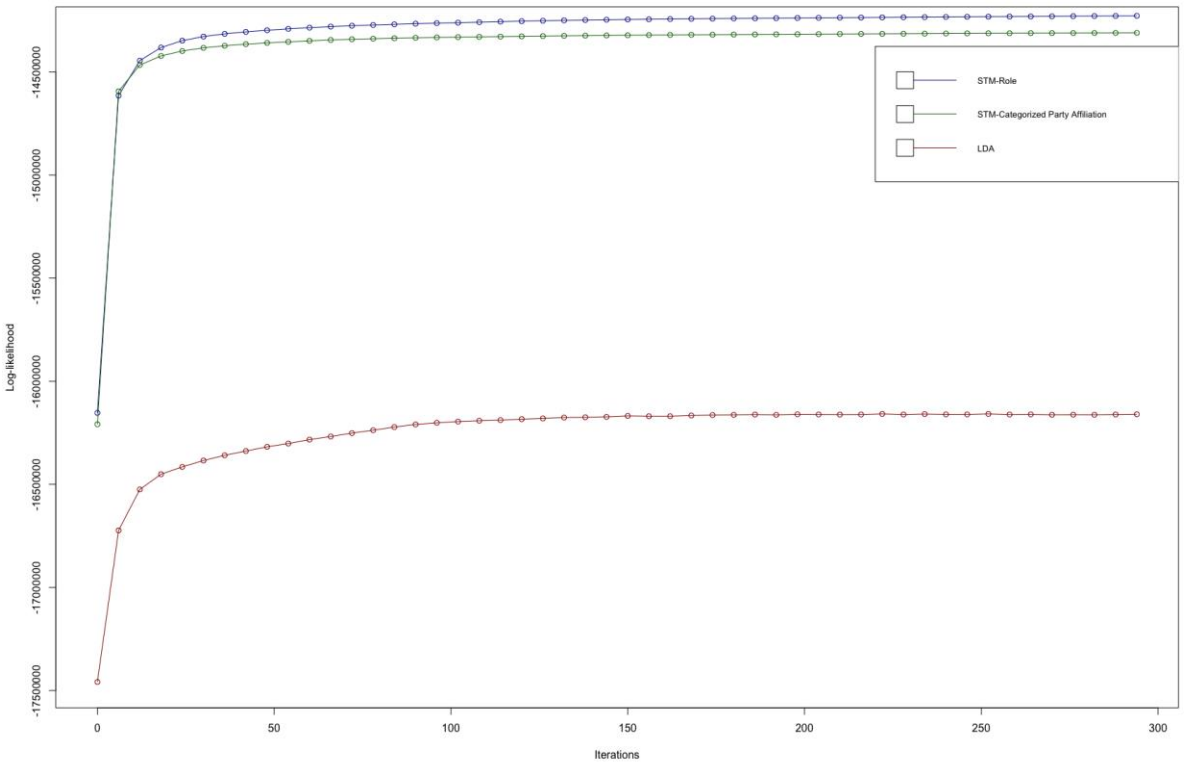


Figure 20: An evaluation of the log-likelihood for LDA and STM models

3.5 Results of Bootstrap Sampling Distribution

As part of the bootstrapping process, each document's topic loading probability (gamma) was mapped to the relevant activity and plotted in figure 21 for LDA and STM. An x-axis showing the topics; a y-axis showing mean gamma proportions faceted with 16 activities, and models were distinguished with red and blue colours. The red line shows the topic proportions of LDA, and the blue lines show the STM. The LDA model showed a high probability for the departments of education and research and rural affairs with 0.57 and a low probability for the departments of gender equality with 0.28 and Finance with 0.25. Furthermore, the STM model also showed a high probability for the departments of education and research and rural affairs with 0.77 and a low probability for the departments of gender equality with 0.23 and Finance with 0.35. For 15 departments, STM had a higher degree of prediction than LDA; only Gender Equality had a slight improvement of 0.05 with LDA.

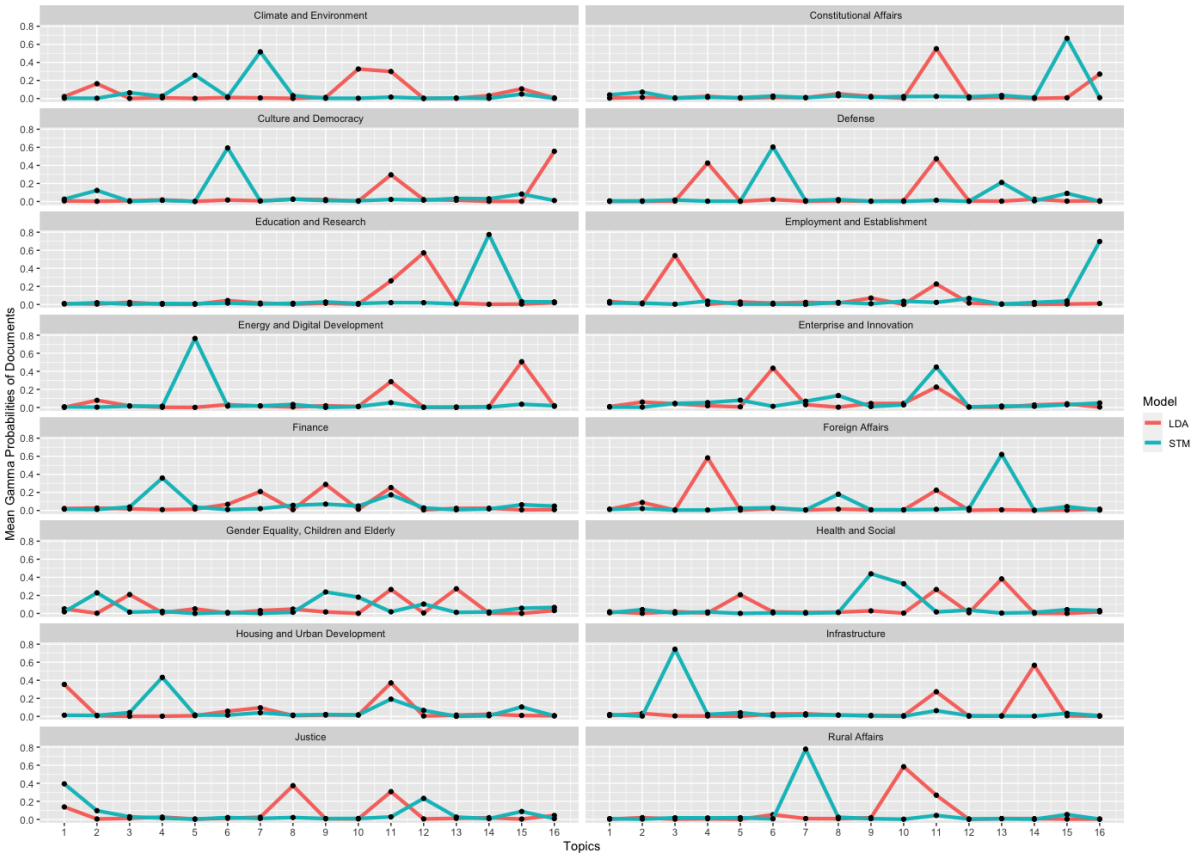


Figure 21: Topic proportions of the ministry activities: According to the fitted models LDA and STM - The proportion of the document relevant to topics was calculated using the mean gamma probabilities posterior distribution

By observing the above findings on the original data, other parameter estimations of the bootstrap samples 200 were calculated for both LDA and STM models and summarized accordingly. Further MeanCI statistics were calculated on the parameter, mean gamma probabilities for both LDA and STM models, and results shown in figure 22. The error bar of confidence intervals of mean gamma probabilities is shown on the x-axis, and activities are shown on the y-axis and faceted with the model category. The documents from a specific debate were compiled extensively loaded on one particular topic in the fitted model. Even though the loadings in most cases were low, the confidence intervals (CI) were tightly indicating that the results were well replicated. As a result of the random sampling, both the LDA and STM mean probabilities increased for the Defense Department and decreased for the Justice Department during the STM.

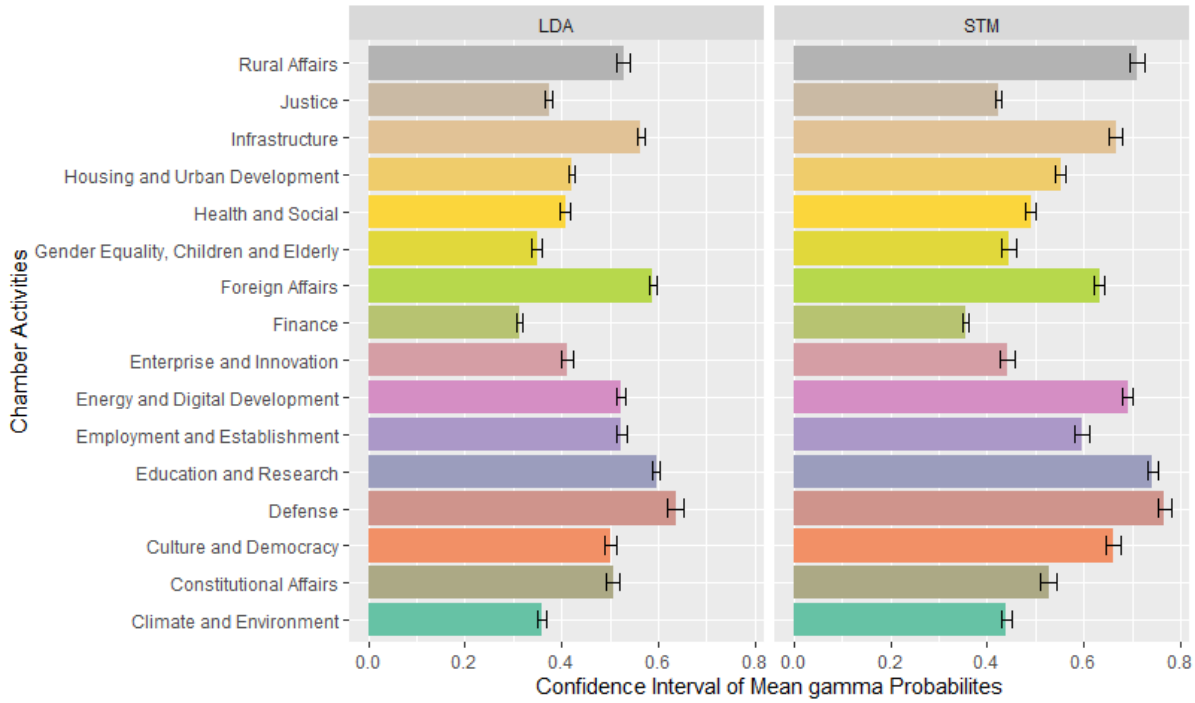


Figure 22: Confidence Intervals of bootstrap parameter for LDA and STM - The mean gamma probabilities were computed using the Topic proportions of each individual activity document of the fitted model

The p-value of the samples t-test was calculated on every single activity by inserting the mean of the original data. The resulting p-value is one (approximately) in all cases where the original data's mean and sampling distribution are the same. In other words, the original results from the original data are well replicated in the bootstrap samples. The resultant metrics assure that the topic modeling is not only useful as the exploratory tool but it can also be used as a confirmatory tool, in that if there are a certain number of underlying (true) topics (ministry of concern in this case), a well formulated topic model is capable of identifying them, with subject to small sampling error. Therefore, in cases where we want to verify whether there is a certain number of specific topics in a corpus, topic modeling can be used there, just like we use confirmatory factor analysis with continuous data.

Chapter 4

Discussion and Conclusion

In our analysis, we investigate the effectiveness of the Swedish parliamentary activities by examining the debate documents. In our research team, there was no native Swedish speaker. Thus, we could not use the original data from the parliament, and we used a secondary source of data that was pre-processed by experts at the Gothenburg University in Sweden.

At first, LDA was implemented to explore the topics according to the type of debate. Here observed that Foreign policy and budget debates could be distinguished as individual topics, and three of the debates, the Special debate, Debates on Parliamentary business, and interpellation debates, discussed the most diverse subjects. Further categorized the documents to connect the analysis according to the ministry of the parliament. Corpus was categorized according to sixteen parliamentary activities, and LDA was able to distinguish seven of the topics, with the Urban Development ministry was showing the quality of highest coherence 0.391 and Employment and Establishment showing the highest prevalence with 7.244. In these activities, LDA successfully extracted most of the document proportions as an individual topic.

Furthermore, the STM model was implemented by giving the party affiliation to investigate whether aiding extra information improves the topic proportions towards activities. Using party affiliation as a covariate, the fitted model performed poorly compared to the LDA with a significant decrement of 1986315(10%) in log-likelihood coefficients. As a result, objective 3 has been accepted for Sweden’s parliament, where the content of speeches is determined by the topic at hand, rather than the speech giver’s identity.

Then we found another critical variable speech giver’s position as a covariate, and the performance of the fitted STM model improved 1980982(14%) compared to LDA. And the fitted model shows that, the predictive performance of the topic proportions got increased among fifteen activities using STM compared to LDA. In contrast, the Gender Equality ministry showed a slight improvement of 0.05 with LDA. The Foreign Affairs ministry topic showed the quality of highest coherence -46.40 and prevalence 8.3. The prevalence of other topics improved compared to LDA as it is evident that documents proportions belonged to activities improved with the STM fitted model. STM can also distinguish seven topics very clearly and other topics with less overlapping than LDA. The overlapping was caused by the model extracting the topics based on the semantic correlation between the documents and drawing inference using the metadata speaker role. It turns out that the effectiveness of the parliamentary system depends mainly on the role of the speaker rather than the party of the speaker. We can interpret this fitted model as a precise way of identifying all 16 topics and categorizing them.

Once the party affiliations had been recoded, the STM model was rerun to see if the model's fit had improved. The results have been improved significantly than the original property and shown log-likelihood value -14306631.88, which is higher than LDA (-16200000) and lower than STM (-14219018.95) with speaker role. As a result of these findings, it became evident that left, center and right-wing parties speak with a distinctive vocabulary.

In Beetham's (2008) view, the system's effectiveness improves once parliament members engage in systematic discussion, resulting in a shared vision of the priorities for parliamentary development. Also, the analysis of Sibanda (2017) explained the use of written questions directed at executive members, which can yield precise information from the administration, secure a commitment, and provide context for the detailed analysis. In addition, parliamentary legislation could be checked through qualitative and quantitative measures such as legislation, oversight, representation, and budgeting (Madhavan & Namita, 2008). However, none of these measures provides information on how coherent the members of parliaments are in their discussion.

According to Magnusson et al.(2018), when the SD party entered the parliament, the concept of immigration and their strategy towards addressing the debates were changed. Our research also found that the strategy of their debate may have influenced other parties. On the contrary, if SD did not influence others, it would speak differently. However, the results indicate that all the parties are not speaking with different vocabulary, though the left-right might have a different voice.

Conducted a simulation study with Bootstrap sampling distribution to assess the uncertainties in the performance of topic modeling with LDA and STM. As a result of the bootstrap analysis, the results of the LDA and STM models can be used for confirmatory analysis since they do not differ significantly over repeated samples.

4.1 Scope of future work

Future research may consider creating a framework with a Finite mixture approach to include metadata and compare it with STM model performance. Through a simulation study, it might be possible to identify and compare the performance of true clusters as part of the evaluation process, using the well-known precision-recall trade-off, and to develop a statistical model balancing the contributions of different types of errors.

Considering that the data used for the research is from a secondary source, it would be beneficial to conduct further research on this issue by political scientists.

Longer time series to study time trends, more background variables, e.g. time length of the speech, main speech or point-of-order, et al., can be incorporated to fit the model better. These models might be tested with speeches from weak parliamentary systems, e.g., countries with weak democratic institutions.

References

Beetham, D., 2008. Evaluating Parliament: A self-assessment toolkit for parliaments (No. 54). *Inter-Parliamentary Union.*.

Blei, D. M., 2012. Probabilistic topic models. *Communications of the ACM, 55(4),* pp. 77-84.

Blei, D. M. & Lafferty, J. D., 2007. A correlated topic model of science.. *The annals of applied statistics, 1(1),* pp. 17-35.

Blei, D. M. & Lafferty, J. D., 2009. *Topic models. In Text mining (pp. 101-124).*. s.l.: Chapman and Hall/CRC..

Blei, D. M., Ng, A. Y. & Jordan, M. I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research, 3,* pp. 993-1022..

Borin, L. et al., 2016, November. Sparv: Språkbanken's corpus annotation pipeline infrastructure.. *In The Sixth Swedish Language Technology Conference (SLTC), Umeå University,* pp. (pp. 17-18)..

Campbell, J. C., Hindle, A. & Stroulia, E., 2015. Latent Dirichlet allocation: extracting topics from software engineering data. *In The art and science of analyzing software data ,* pp. 139-159.

Campbell, J. E., 1983. Ambiguity in the issue positions of presidential candidates: A causal analysis. *American Journal of Political Science,* pp. 284-293.

Danielsson, J., de Haan, L., Peng, L. & de Vries, C. G., 2001. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis, 76(2),,* pp. 226-248.

DiCiccio, T. J. & Efron, B., 1996. Bootstrap confidence intervals. *Statistical science, 11(3),* pp. 189-228.

Eichorst, J. & Lin, N. C., 2019. Resist to commit: Concrete campaign statements and the need to clarify a partisan reputation. *The Journal of Politics, 81(1),* pp. 15-32.

Eide, S. R., 2020. "Anföranden: Annotated and Augmented Parliamentary Debates from Sweden". *Proceedings of the Second ParlaCLARIN Workshop,* pp. 5-10.

Eisenstein, J., Ahmed, A. & Xing, E. P., 2011. Sparse additive generative models of text. *In Proceedings of the 28th international conference on machine learning (ICML-11),* pp. pp. 1041-1048.

Government, S., 2021. [Online]
Available at: https://www.government.se/government-of-sweden/
[Accessed 2021].

Greene, D. & Cross, J. P., 2015, June. Unveiling the political agenda of the European Parliament plenary: A topical analysis. *In Proceedings of the ACM web science conference (pp. 1-10).*

Grimmer, J., 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18(1),* pp. 1-35.

Grimmer, J. & Stewart, B. M., 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis, 21(3),* pp. 267-297.

Harden, J. J., 2011. A bootstrap method for conducting statistical inference with clustered data. *State Politics & Policy Quarterly, 11(2),* pp. 223-246.

Hausser, J. & Strimmer, K., 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks.. *Journal of Machine Learning Research,* p. 1469–1484..

Hofmann, T., 1999, August. Probabilistic latent semantic indexing. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 50-57.

Hopkins, D. J. & King, G., 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54(1) ,* pp. 229-247.

Hornik, K. & Grün, B., 2011. topicmodels: An R package for fitting topic models. *Journal of statistical software, 40(13),* pp. 1-30.

Jelodar, H. et al., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78(11),* pp. 15169-15211.

Jones, T., Doane, W. & Jones, M. T., 2016. Package 'textmineR'. Functions for Text Mining and Topic Modeling..

Kumawat, D. & Jain, V., 2015. POS tagging approaches: a comparison. *International Journal of Computer Applications,* p. 118(6).

Kwartler, T., 2017. Text mining in practice with R. John Wiley & Sons..

Lapponi, E., Søyland, M. G., Velldal, E. & Oepen, S., 2018. The Talk of Norway: a richly annotated corpus of the Norwegian parliament,1998–2016. *Language Resources and Evaluation, 52(3),* pp. 873-893.

Lee, M. & Mimno, D., 2014. Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference.. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),,* p. 1319–1328.

Madhavan & Namita, 2008. Measuring the Effectiveness of the Indian Parliament Background. *PRS Legislative Research Centre for Policy Research.*

Magnusson, M., Öhrvall, R., Barrling, K. & Mimno, D., 2018. Voices from the far right: a text analysis of Swedish parliamentary debates.

Meguid, B. M., 2005. Competition between unequals: The role of mainstream party strategy in niche party success.. *American Political Science Review, 99(3),* pp. 347-359.

Mei, Q., Xuehua, S. & ChengXiang, Z., 2007. Automatic labeling of multinomial topic models. *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* p. 490–499.
Mimno, D. M. & McCallum, A., 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *In UAI,* pp. Vol. 24, pp. 411-418.
Nanni, F., Menini, S., Tonelli, S. & Ponzetto, S. P., 2019, June. Semantifying the UK hansard (1918-2018).. *In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL),* pp. pp. 412-413. IEEE.
Novak, P., 2018-2021. *XiMpLe XML Editor.* [Online]
Available at: http://www.ximple.cz/index.php
Purba, N. S. & Nooraeni, R., 2019, August. *Using LDA for Innovation Topic of Technology: Quantum Dots Patent Analysis.* s.l., s.n.
Python Documentation, 2001-2022. *Python Software Foundation.* [Online]
Available at: https://docs.python.org/3/library/xml.etree.elementtree.html
[Accessed 2021].
Qiaozhu, M., Xuehua, S. & ChengXiang, Z., 2007. Automatic labeling of multinomial topic models. *In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* p. 490–499.
Quinn, K. M. et al., 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54(1) ,* pp. 209-228.
RIKSDAG, S., 2021. [Online]
Available at: https://www.riksdagen.se/en/how-the-riksdag-works/the-work-of-the-riksdag/debates-and-decisions-in-the-chamber/
[Accessed 2021].
Roberts, M. E., Stewart, B. M. & Airoldi, E. M., 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111(515),* pp. 988-1003.
Roberts, M. E., Stewart, B. M. & Tingley, D., 2019. Stm: An R package for structural topic models. *Journal of Statistical Software, 91(1),* pp. 1-40.
Rortais, A. et al., 2021. A topic model approach to identify and track emerging risks from beeswax adulteration in the media. *Food Control, 119, 107435..*
Sibanda, R., 2017. Assessing the effectiveness of written questions & Replies as an oversight mechanism.
Slapin, J. B. & Proksch, S. O., 2008. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52(3),* pp. 705-722.
spraakbanken, 2021. *spraakbanken.* [Online]
Available at: https://spraakbanken.gu.se/korp/markup/msdtags.html
[Accessed 2021].
SpråkbankenText, 2018. [Online]
Available at: https://spraakbanken.gu.se/en/resources/rd-anf
[Accessed 2021].
T.Hofmann, 1999. Probabilistic latent semantic indexing. *in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,* pp. 50-57.
Tong, Z. & Zhang, H., 2016, May. A text mining research based on LDA topic modelling. *In International Conference on Computer Science, Engineering and Information Technology ,* pp. 201-210.
Wesslen, R., 2018. Computer-assisted text analysis for social science: Topic models and beyond. *arXiv preprint arXiv:,* p. 1803.11045.
Zivot, E., 2017. *Introduction to computational finance and financial econometrics..* s.l.:Chapman & Hall Crc..
GithubRepo: https://github.com/vinuvish/MicrodataAnalysis_final_thesis

Appendix A
Sweden Government ministries

Table A: Categorization of Sweden Government ministries (Period: 2014-2018)

Category of Ministry	Ministry Responsibilities	Minister Name	Serving Period
Prime Minister Minister for EU Affairs	Leading and coordinating the work of government offices	Stefen Löfven	2014-current
Ministry of Culture	Culture, Democracy and Human rights, Media, Sport, Youth policy, Civil society	Alice Bah Kuhnke (MP)	2014-2019
Ministry of Defense	Military operations and its supporting agencies	Peter Hultqvist (S)	2014-Current
Ministry of Education and Research	School Performance, conditions for teachers and Study financing	**Minister of Education:** Gustav Fridolin (MP) **Minister for Upper Secondary School, Adult Education and Training:** Aida Hadzialic (S) Anna Ekström (S) **Minister for Higher Education & Research:** Helene Hellmark Knutsson (S)	2014-2019 2014-2016 2016-2019 2014-2019
Ministry of Employment	Labour market, Labour law and work Environment, Gender Equality, Human rights at National Level, Integration Efforts, and Children's Rights, Urban development, anti-segregation, and anti-discrimination	**Minister of Employment:** Ylva Johansson (S) **Minister for Gender Equality, Childern and Elderly**: Åsa Regnér (S) Lena Hallengren (S) **Minister for Housing and Urban Development:** Mehmet Kaplan (MP) Peter Eriksson (MP)	2014-2019 2014-2018 2018-2019 2014-2016 2016-2019
Ministy for Enterprise and Innovation	Enterprise and Industrial policy, State-owned Enterprises, Innovation, Regional Development, Rural Affairs	**Minister for Enterprise:** Mikael Damberg (S) **Minister for Rural Affairs:** Sven-Erik(S)	2014-2019 2014-2019
Ministry of Environment	Environment, Climate, Chemicals, Nature Conservation, Forest Conservation, Marine Environments, Water Environments, and Radiation Safety	Åsa Romson (MP) Karolina Skog (MP)	2014-2016 2016-2019
Ministry of Finance	Central Government Administration, Consumer Legislation, Economic Policy, Financial markets, Housing and community planning,	**Minister for Finance:** Magdalena Andersson (S)	2014-2021 2014-2019

	Municipalities and regions, Public Procurement, Taxes, and tariffs	**Minister for Public Administration:** Ardalan Shekarabi (S) Minister of Financial Marketing and Deputy minister for Finance: Per Bolund MP	2014 - 2021
Ministry for Foreign Affairs	Democracy and human rights, Foreign and security policy, International development cooperation, Trade, and promotion of Sweden	**Minister for Foreign Affairs:** Margot Wallström (S) **Minister for Foreign Trade and Nordic Affairs:** Kristina Persson (S) Anna Linde(S) Margot Wallström (S) **Minister for International Development Cooperation:** Minister Isabella Lövin (MP)	2014-2019 2014-2016 2016-2019 2016-2019 2014-2019
Ministry of Infrastructure	Digital Policy, Energy, Post Issues, Transport, and Infrastructure	**Minister for Infrastructure:** Anna Johansson (S) Tomas Eneroth (S) **Minister for Energy and Digital Development:** Ibrahim Baylan (S)	2014-2017 2017-Current 2014-2019
Ministry of Health and Social Affairs	Disabilities, Public health and medical care, social insurance, Social services including care for older people	**Minister for Health and Social Affairs:** Gabriel Wikström (S) **Minister for Social Security:** Annika Strandhäll (S)	2014-2017 2014-2019
Ministry of Justice	Judicial system, The Constitution of Sweden and personal privacy, Combating terrorism, Emergency preparedness, Family law, Migration, and asylum	**Minister for Justice and Migration:** Morgan Johanssonn (S) **Minister for Migration and Aylum:** Heléne Fritzon (S) **Minister for Home Affairs or Minister for Interior:** Anders Ygeman (S) Morgan Johansson(s)	2014-2017 2017-2019 2014-2017 2017-2019

Appendix B
Parliamentary Debates Speech Giver's Position

Table B: Categorization of Speech Giver's Position

Serial Number	Position
1	Vice gruppledare
2	Finance Minister
3	Ledamot
4	Suppleant
5	Vice ordförande
6	Riksdagsledamot
7	Justice and Migration Minister
8	Ersättare
9	Justice Minister
10	Ordförande
11	Förste vice ordförande
12	Health and Social Affairs Minister
13	Statsrådsersättare
14	Rural Affairs Minister
15	Culture Minister
16	Education Minister
17	Infrastructure Minister
18	Foreign Affairs Minister
19	Employment Minister
20	Landsbygdsminister
21	Defense Minister
22	Innovation Minister
23	Social Affairs Minister
24	Gruppledare
25	Deputerad
26	Talmansersättare
27	Personlig suppleant
28	Partiledare
29	Labor Minister
30	Extra suppleant
31	Climate and Environment Minister

Appendix C
Quality metrics of the fitted models

Table C1: Listing values of quality factors (Coherence, Prevalence) of top terms and their respective labels for LDA model

Topic	Coherence	Prevalence	Top Terms	Classified Labels
Topic 1	0.311	3.123	bank, finansiell, skuld, avgift, lån	Housing and urban development, Finance, Justice
Topic 2	0.291	7.244	arbetsmarknadsminister, upphandiling, kollektivavtal, arbetslös, arbetsmarknadspolitik	Employment and Establishment, Finance, Gender Equality ministries
Topic 3	0.054	6.580	granskning, böra, granska, statsminister, behandla	Constitutional affairs
Topic 4	0.105	6.025	vinst, handel, företagande, industri, starta	Enterprise and innovation, Finance, Foreign affairs
Topic 5	0.227	4.936	gymnasieskola, högskola, undervisning, huvudman, gymnasium	Education and Research
Topic 6	0.335	6.183	humanitär, fred, krig, utrikespolitik, stödja	Foreign Affairs
Topic 7	0.391	3.388	bostadsbyggande, byggande, bostadsmarknad, hyresrätt, bostadspolitik	Housing and Urban development
Topic 8	0.386	4.939	brottslighet, straff, grov, begå, asylsökande	Justice
Topic 9	0.164	3.241	Avdrag, pension, pensionär, inkomst, skattesänkning	Finance, Health and Social
Topic 10	0.225	4.838	fordon, transport, köra, flyg, lastbil	Infrastructure
Topic 11	0.317	3.657	public service, museum, film, kulturarv, kulturminister	Culture and Democracy
Topic 12	0.292	6.739	försvarsmakt, försvarsminister, nordisk, military, Nato	Foreign affairs, Defense
Topic 13	0.076	27.492	avse, föila, tacka, fatta, tack	Contains unclassified adjectives
Topic 14	0.235	3.491	förnybar, energi, kärnkraft, klimat, industri	Energy and Digital development, Climate and Environment
Topic 15	0.237	4.072	psykisk, läkemedel, ohälsa, tillgänglighet, sjukdom	Health and Social
Topic 16	0.292	5.482	djur, jordbruk, varg, jakt, produktion	Rural affairs, Climate and Environment

Table C2: Listing values of quality factors (Semantic Coherence, Prevalence) of top terms and their respective labels for STM model

Topic	Semantic Coherence	Prevalence	Top Terms	Classified Labels
Topic 1	-62.26199	6.243463	straffrabatt, vapenbrott, narkotikabrott, polisutbildning, stöldliga, bostadsinbrott	Justice
Topic 2	-87.77184	3.809105	våld, public service, flicka, sexuell, utsätta, diskriminering, barnäktenskap	Gender Equality, Children and Elderly
Topic 3	-54.55335	7.308784	trafiksäkerhet, alkobom, cyklist, farledsavgift, infrastrukturproposition, cykelstrategi,	Infrastructure, Gender Equality
Topic 4	-59.35774	8.110893	marginalskatt, amorteringskrav, skatteutgift, penningpolitik, hyresrätt, bostadspolitik,	Finance,
Topic 5	-51.47254	4.721989	elproduktion, reaktor, uran, effektskatt, elpris, elcertifikatssystem, uranbrytning, effektreserv, energisystem,	Housing and Urban Development
Topic 6	-100.95907	5.583830	filmpolitik, kulturpolitik, biograf, kulturpolitisk, krigsförband, kulturskapare, filmavtal, museum, kulturminister,	Culture and Democracy
Topic 7	-52.33401	7.050160	skog, djur, jordbruk, varg, jakt, produktion, biologisk	Climate and Environment
Topic 8	-73.06407	4.713730	arbetsprogram, handelsavtal, handelspolitik, frihandel, frihandelsavtal, gränshinder,	Enterprise and Innovation
Topic 9	-68.99356	6.267220	patientsäkerhet, sjukvårdslag, ungdomspsykiatri, patient, patientlag	Health and Social
Topic 10	-68.92025	4.991188	pension, jämställdhet, pensionär, sjukförsäkring, föräldraförsäkring, sjuk, inkomst	Gender Equality, Children and Elderly, Health and Social
Topic 11	-74.88278	7.029913	bolag, upphandling, konsument, bank, företagande, sälja, konkurrens	Culture and Democracy

Topic 12	-46.40581	5.380253	asylsökande, uppehållstillstånd, tillfällig, mottagande, ensamkommande, asyl, flykting	Foreign affairs
Topic 13	-51.14897	8.379167	utrikesminister, bistånd, militär, humanitär, Nato, fred, krig	Foreign Affairs
Topic 14	-61.10130	6.824655	förskoleklass, lärarutbildning, årskurs, undervisningstid, Skolinspektion, yrkesprogrammen, skollag,	Education and Research
Topic 15	-80.46430	7.400019	riksrevisor, regeringsform, konstitutionell, grundlag, presstöd, nämnd, granskning, valsedel, offentlighet, utfrågning,	Rural affairs
Topic 16	-57.39212	6.185630	extratjänst, arbetsmarknadsutbildning, nystartsjobb, arbetsmarknadsminister, traineejobb,	Employment and Establishment