UPPSALA
UNIVERSITET

UPTEC STS 20037
Examensarbete 30 hp
December 2020

Exploring NMF and LDA Topic Models of Swedish News Articles

Johan Blad
Karin Svensson

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Exploring NMF and LDA Topic Models of Swedish News Articles

Johan Blad, Karin Svensson

The ability to automatically analyze and segment news articles by their content is a growing research field. This thesis explores the unsupervised machine learning method topic modeling applied on Swedish news articles for generating topics to describe and segment articles. Specifically, the algorithms non-negative matrix factorization (NMF) and the latent Dirichlet allocation (LDA) are implemented and evaluated. Their usefulness in the news media industry is assessed by its ability to serve as a uniform categorization framework for news articles. This thesis fills a research gap by studying the application of topic modeling on Swedish news articles and contributes by showing that this can yield meaningful results. It is shown that Swedish text data requires extensive data preparation for successful topic models and that nouns exclusively and especially common nouns are the most suitable words to use. Furthermore, the results show that both NMF and LDA are valuable as content analysis tools and categorization frameworks, but they have different characteristics, hence optimal for different use cases. Lastly, the conclusion is that topic models have issues since they can generate unreliable topics that could be misleading for news consumers, but that they nonetheless can be powerful methods for analyzing and segmenting articles efficiently on a grand scale by organizations internally. The thesis project is a collaboration with one of Sweden's largest media groups and its results have led to a topic modeling implementation for large-scale content analysis to gain insight into readers' interests.

Handledare: Lovisa Bergström
Ämnesgranskare: Niklas Wahlström
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 20037

Populärvetenskaplig sammanfattning

Nyhetsmedier spelar en fundamental roll för fungerande demokratiska samhällen. Ett av deras främsta sätt att kommunicera är genom text, exempelvis i form av nyhetsartiklar. Den digitala revolution vi just nu lever i möjliggör för dessa samhällsaktörer att sprida texter till fler och snabbare genom att digitalisera och att lagra texter som till exempel digitala nyhetsartiklar, poster på sociala medier och liknande. Denna utveckling sker i en allt högre grad vilket skapar enorma digitala samlingar av nyhetsartiklar, som försvårar att manuellt analysera sådana samlingar av artiklar. Detta faktum har öppnat upp ett nytt, spännande och högaktuellt forskningsområde, nämligen möjligheten att automatiskt analysera och segmentera nyhetstexter. Bakgrunden till detta behov är att analys av nyhetstexter ger en ökad förståelse för vad människor läser och bryr sig om, eftersom sådana intressanta mönster går att härleda från dessa textsamlingar.

En ytterligare anledning till att detta forskningsområde får mycket publicitet är att tekniker under namnet maskininlärning tar en allt större roll, både i akademin, men också i samhället generellt. Maskininlärning kan kortfattat beskrivas som den vetenskapliga studien av automatiska metoder för dataanalys. Maskininlärningsmetoder kan automatiskt upptäcka mönster i data och använda dessa mönster för att förutspå framtida resultat utifrån data eller utföra andra typer av uppgifter, exempelvis att gruppera data. Maskininlärning delas traditionellt upp i två fack, övervakad och oövervakad maskininlärning. I övervakad maskininlärning förser man metoden med märkt data, det vill säga data som innehåller exempel på önskade svar. I oövervakad maskininlärning förser man metoden med omärkt data utan någon tydlig struktur, och låter metoden själv hitta om det existerar några relationer mellan data. Topic modeling, direkt översatt ämnesmodellering, är en oövervakad maskininlärningsmetod vars syfte är att hitta ämnen ur textsamlingar genom att analysera orden i texterna. Detta görs genom att använda statistiska relationer mellan ord för att ta fram ämnen som utgörs av ord med hög sannolikhet att beskriva artiklar inom ämnet. Vanligtvis väljs sedan de fem eller tio ord med högst sannolikhet som själva ämnet. Ett ämne kan således exempelvis vara: *thunberg greta klimataktivist klimat värld*. Ämnesmodellering har i tidigare forskning flertalet gånger applicerats på nyhetsartiklar, men då i huvudsak på engelska nyhetsartiklar. Detta lämnar frågan kring hur välfungerande metoden är på svenska nyhetsartiklar obesvarad.

Denna studie ämnar besvara ovanstående fråga genom att utforska hur ämnesmodellering fungerar med svenska nyhetsartiklar. För att möjliggöra detta genomförs studien i samarbete med Bonnier News, ett av Sveriges ledande mediehus som når över tre miljoner användare varje dag. Mer specifikt undersöks ämnesmodellerings förmåga att generera meningsfulla ämnen för att beskriva och segmentera nyhetsartiklar, och på det sättet användas som grund för ett kategoriseringsramverk för nyhetsartiklar. Det finns flera olika ämnesmodelleringsmetoder varav de två mest aktuella i dagens forskningsläge utvärderas i denna studie, nämligen icke-negativ matrisfaktorisering (NMF) och latent Dirichlet allokering (LDA). NMF är

en metod med teoretisk grund i linjär algebra och LDA har sin teoretiska bakgrund i sannolikhetslära. I denna studie applicerades de två metoderna på nyhetsartiklar från Bonnier News, för att sedan kvantitativt och kvalitativt utvärdera de genererade ämnena, samt jämföra mot manuella sätt att analysera och segmentera nyhetsartiklar. Den kvantitativa utvärderingen gjordes huvudsakligen genom att mäta samhörighet mellan de tio termer med högst sannolikhet för ämnet, medan den kvalitativa utvärderingar utfördes genom att observera semantiken för de ämnen som metoderna genererar. Metoderna applicerades på data från artiklar från ett enskilt nyhetsmedium (i detta fall Dagens Nyheter), men även på data från flera olika nyhetsmedier för att undersöka hur metoderna påverkas av skillnader i textdata mellan olika nyhetsmedier.

En stor del av studien innefattar att hitta de bästa sätten att förbehandla textdata som är avsedd att användas med ämnesmodellering. Förbehandling syftar här till att representera text i en form som kan tolkas av datorer, för att sedan kunna användas av maskininlärningsmetoder. Detta kan exempelvis innebära att beräkna hur vanligt ett visst ord är i en artikel, eller att filtrera ut vissa ordklasser. Alla typer av automatiska metoder applicerade på text kräver olika typer av förbehandling för att nå bra resultat och ämnesmodellering är inget undantag.

Resultaten i studien visar på att ämnesmodellering med svenska nyhetsartiklar genererar meningsfulla ämnen. Dock krävs omfattande förbehandling av nyhetsartiklarna, exempelvis genom att filtrera ut ordklasser. I synnerhet skapas meningsfulla ämnen då man filtrerar ut alla ordklasser förutom substantiv, och studiens resultat visade på att det kan vara fördelaktigt att även filtrera ut namngivna substantiv (namn, länder, städer etc.). Vidare visas att både NMF och LDA ger framgångsrika resultat. Vilken modell som är att föredra är högst beroende av i vilket syfte modellen ska användas då båda metoderna har fördelar och nackdelar. Det som dock kan konstateras är att ämnesmodellering har tydliga problem, då metoden i vissa fall skapar opålitliga ämnen som kan vara vilseledande för en nyhetskonsument. De genererade ämnena bör därav inte presenteras för en slutanvändare. Ämnesmodellering är däremot en kraftfull metod för att effektivt analysera och segmentera artiklar i stor skala av nyhetsmedier för interna analyser, och har därmed en stor fördel jämfört med manuella metoder. Utifrån resultaten från denna studie har en ämnesmodellerings-metod implementerats hos Bonnier News, som kommer att användas för att genomföra interna innehållsanalyser av nyhetsartiklar samt för att få en ökad förståelse för deras användares läsarmönster.

Acknowledgments

This thesis is the result of a Master's Thesis project by Johan Blad and Karin Svensson which was conducted together with Bonnier News and the Department of Information Technology at Uppsala University. We would like to thank our supervisors, Lovisa Bergström and Abtin Salahshor from the data analytics team at Dagens Nyheter (part of Bonnier News), for all their valuable support and help along the way. Many thanks also to the machine learning team at Bonnier News for their technical expertise. Finally, we would like to thank our university subject reader Niklas Wahlström for his valuable insights and guidance.

Johan Blad and Karin Svensson
December 2020

Distribution of Work

This thesis has been written by Johan Blad and Karin Svensson who have collaborated on all areas covered in this thesis. Both authors have been equally responsible for project planning, management, and communication with stakeholders at Bonnier News for the entire master thesis project. In general, Johan was responsible for the NMF algorithm and data preparation, hence wrote most of the code related to those areas as well as the corresponding parts throughout the thesis. He also wrote the chapter on related work and most of the code for the implementation of the model in Bonnier News IT infrastructure. Karin was responsible for the LDA algorithm and data management, hence wrote most of the code related to those areas as well as the corresponding parts throughout the thesis. She also wrote most of the code for the visualizations of the model results for the thesis and the visualization reports for the model results in Bonnier News IT infrastructure. The remaining chapters of the thesis were written by both authors by an equal workload.

Abbreviations

ADJ	Adjectives
BN	Bonnier News
CD	Coordinate descent
CNOUN	Common nouns
Di	Dagens Industri
DN	Dagens Nyheter
HD	Helsingborgs Dagblad
LDA	Latent Dirichlet allocation
LSA	Latent semantic analysis
LSI	Latent semantic indexing
MCMC	Markov chain Monte Carlo
MU	Multiplicative update
NER	Named entity recognition
NLP	Natural language processing
NMF	Non-negative matrix factorization
NNDSVD	Nonnegative double singular value decomposition
NNDSVDa	Nonnegative double singular value decomposition with the zero elements as the average
pLSA	Probabilistic latent semantic analysis
pLSI	Probabilistic latent semantic indexing
POS	Part of speech
PROPN	Proper nouns
SVD	Singular value decomposition
tf-idf	Term frequency-inverse document frequency
VB	Variational Bayes

Table of Contents

1 Introduction

Text data has been one of the most important sources of information and knowledge for humanity throughout modern history, especially so today. The ability to derive insights and understandings from such data is therefore a great enabler of successful knowledge dissemination in society overall. Journalism and the news media industry in general is a key component in this process, and the majority of their content is presented through text. These societal actors play a crucial role with a high impact on what information reaches out to the general public and they are fundamental to the functioning of a democratic society. The digital revolution of modern times has tremendously enhanced the possibilities for media organizations to spread their content and widen their reach. At the same time, as sources of information and knowledge from these media continue to be digitized and stored in the form of news articles, webpages, social media posts, and alike, it becomes more and more difficult to derive insights from these enormous sources of data.

This is where automated methods to extract information from such data come into play, an area of research in the field of machine learning that has become increasingly popular (Grimmer and Stewart 2013). One prominent method that takes advantage of enormous amounts of text data and explores it is the unsupervised machine learning method topic modeling. This method aims to discover themes that run through a collection of documents by using statistical relationships between the terms in these documents. Topic modeling algorithms do not require any prior annotations or labeling of the documents, meaning they can analyze the content of large amounts of text without requiring manually coded labels, thus reducing the time and costs of such projects. Topic models instead generate topics from these texts and assign documents to them, where a topic consists of terms that are statistically related in the document collection (Jacobi, Van Atteveldt, and Welbers 2016; Blei 2012).

Topic modeling has broad applications in various contexts of which news articles are one widely researched area. Analysis of media content is widely acknowledged as a way to understand what information is received by the public and how it is framed, but it is a tedious process when done manually (Chandelier et al. 2018; Q. Liu et al. 2019). Topic modeling is a tremendously more time-efficient substitute that can yield valuable insights in this data as general patterns in large article collections tell a story of what people read and care about (Surjandari et al. 2018). Topic modeling also has the benefit that it generates topics independently from human preconceptions, and can potentially lead to unexpected but valuable results such as hidden patterns and relationships between articles. This has sparked many value creation use cases. For example, topic modeling has been used as a tool for investigating trends and to analyzing variations in media content such as coverage of specific issues (Chandelier et al. 2018), or as a uniform framework for categorization of articles without human labeling (Surjandari et al. 2018). The inherent value of topic modeling for all of its use cases stems from its potential

to reduce vast data sources into meaningful topics, where meaningful alludes to interpretable and useful results to the practitioners that make use of them.

Most previous studies within this research field have been made with English news articles, leaving a research gap in the literature on how topic models perform on articles in other languages. This research aims to partially fill that gap by applying topic models to Swedish news articles, in collaboration with the organization Bonnier News (BN). BN is one of Sweden's leading news media groups with a reach of 3 million readers each day (Bonnier News 2020). The organization had the incentive to better understand their content in terms of what major themes exist in their news articles and how different articles relate to one another. The BN group covers many brands, spanning from lifestyle magazines, industry magazines, local newspapers to Sweden's biggest nationwide morning newspaper. This variety gives possibilities for cross-brand activities such as cross-brand marketing but has also led to problems such as lack of a uniform framework for categorization of articles.

Dagens Nyheter (DN) is one of BN's biggest brands and Sweden's biggest nationwide morning newspaper. DN currently categorizes their articles manually, with sections and tags. The journalists define what section an article belongs to, for example, "Ekonomi" (Economy), and add a few tags, for example, "BNP" (gross national product) and "Finansminister" (Minister of Finance). Expressen is another brand within BN, having other names for their sections and tags, as well as Dagens Industri, a third brand with other sections and tags. This list continues, leaving Bonnier News with a variety of sections and tags across their brands. However, all brands share the common denominator of having news articles that are about one or several underlying topics. An automated method that creates a uniform categorization framework, independently from human preconceptions, for these articles across all brands at BN could enable new ways for value creation. By utilizing the vast data sources of articles at BN, this research aims to evaluate how topic models best can be used to create value in content analysis and categorization tasks in the scientific field of automated media analysis in a Swedish context.

1.1 Research Aim

This thesis explores the application of automated topic generation by the unsupervised machine learning method topic modeling of Swedish news article text data. Different such models were applied to news article datasets of different news brands and evaluated by quantitative metrics and qualitative human judgment, to investigate the potential of topic modeling of Swedish news articles. Further, the aim is to evaluate if topic modeling can be used to develop a uniform categorization framework for Swedish news articles. Supported by this context, the purpose of this thesis is to provide an understanding of the validity, the viability of use, and limitations for such a topic model categorization framework for news articles from one brand or spanning multiple brands.

To assess the research purpose, the following research questions are posed:

1. Which topic model algorithm and data preparation yields the most meaningful topics of Swedish news articles within one news brand?
2. How well do topic model algorithms generalize to Swedish news article datasets of multiple news brands?
3. How do topic modeling and the created topics differ from other categorization methods for Swedish news articles, and what are the strengths and weaknesses compared to these?
4. Is topic modeling a viable framework for categorization of Swedish news articles from multiple brands and what are its main advantages and limitations?

1.2 Disposition

The thesis is divided into nine chapters. Chapter two presents the theoretical background and serves as a basis for the concepts that are used throughout the thesis. Chapter three covers related work, giving the thesis context by presenting previous research in the area of topic modeling and other techniques with a similar purpose. In chapter four, the data which was used is presented. Chapter five presents the method and its underlying methodology for building and evaluating meaningful topic models. In chapter six the results from the experiments are presented and this is followed by a discussion of the results in chapter seven. Chapter eight consists of the conclusive findings and proposed future research. Finally, chapter nine covers a topic model implementation at Bonnier News and the value that it can provide.

1.3 Delimitations

The implemented and evaluated topic modeling methods in this thesis are delimited to two current state-of-the-art algorithms, namely non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA). Other topic modeling algorithms do exist but judging from previous research NMF and LDA are generally acknowledged to be the best-suited algorithms for topic modeling of text documents which motivates this delimitation. Furthermore, data preparation is a key component of topic modeling applications but there are many different methods that could be utilized in this process. Therefore, the choice of data preparation methods used in this thesis is a delimitation, as numerous other techniques could be utilized to enhance topic modeling. The qualitative evaluation of the topic models by human judgment is also a delimitation since a large-scale human evaluation is beyond the scope of this thesis.

The Bonnier News media group and the brands that they cover are in this thesis seen as representative of Swedish news media in general, as the media group have morning papers, evening papers, industry magazines, and other niche brands. However, Sweden does have many other brands with unique content and different categorization methods of their articles. This is a delimitation since Bonnier News does not span the entire Swedish news media field.

2 Theory

This chapter provides a theoretical background and serves as a basis for the concepts that are used throughout this thesis. First, an overview of the chapter will be given, followed by a presentation of the field of machine learning and related important concepts. Then a presentation will follow on methods from the field of natural language processing and lastly a presentation of the theory behind topic modeling.

2.1 Overview

This chapter aims to explain the theory behind unsupervised machine learning in general and topic modeling in particular. An overview is given of the research field of machine learning with a focus on unsupervised machine learning, to give the reader a context for topic modeling.

Topic modeling uses natural language as input. Natural language needs to be preprocessed into a structured format suitable for computations before the actual modeling. Before moving on to the section that explains the algorithms that topic models are built on, some key concepts on how to represent natural language in this suitable format are explained. In literature, these concepts are often referred to as methods in the field of natural language processing (NLP).

Next, the key idea behind topic modeling is presented. In this thesis, two topic modeling algorithms are evaluated, non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA). These algorithms both use text as raw input and give similarly structured output, but differ in calculations on how to derive that output. To understand the difference between these two approaches, a background of how the different algorithms emerged will be given, followed by detailed descriptions of the two algorithms.

Finally, theories on how to evaluate topic models are presented. Two approaches for measuring the quality of a topic model, quantitative and qualitative, will be explored.

2.2 Machine Learning

Machine learning can be described as the scientific study of automated methods for data analysis. In particular, Murphy (2012) defines the area "as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty". Mohri, Rostamizadeh, and Talwalkar (2018) phrase its definition as "machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions". Here experience refers to known information, such as collected data. Both definitions mention predictions, which is the core of machine learning. The basis for making predictions in machine learning are mathematical models, using probability theory. This makes the discipline

of machine learning closely related to the field of statistics but differs in terms of terminology and applications. In essence, the decisions that machine learning models make are based on statistics, where the statistical models take massive amounts of data as input (Murphy 2012).

Machine learning is typically divided into two sub-groups: supervised machine learning and unsupervised machine learning. Supervised machine learning is normally defined as the classic machine learning method, where the objective is to learn a model that maps from input to output, given a labeled set of input-output pairs called a training set. The input is typically observations and the output is the corresponding response variables. The model is learned using the training data, and the subsequent trained model can be used to make predictions for unseen data. This is the most common scenario associated with classification, regression, and ranking problems (Mohri, Rostamizadeh, and Talwalkar 2018).

The other category is unsupervised machine learning. In unsupervised machine learning, the model is learned with unlabeled data. This means that only observations are present, without response variables. The result is that predictions cannot be made as for a supervised machine learning problem, and the goal is instead to make interesting discoveries and find patterns among the observations. This is sometimes called knowledge discovery (Murphy 2012). Examples of unsupervised learning are discovering groups within the data, also known as clustering, to determine the probability distribution of the input data, or to visualize data in lower dimensions by projecting the data from high-dimensional space (Bishop 2006).

Unsupervised learning is arguably more applicable than supervised machine learning since it does not require a labeled data set. It can however be problematic to measure the performance of an unsupervised model since the problem is much less well-defined and the learning algorithm is not told what patterns to look for. Because of this, there is no obvious error metric to use, compared to supervised learning where it is possible to test a model by using labeled data and compare the prediction of a variable with the true observed value (Murphy 2012).

To clarify concepts and reasoning in this thesis, the following machine learning relationships between an algorithm, a model and data are stated. Given a dataset used for the learning process, a machine learning algorithm is applied on that dataset to learn a model. A machine learning model is thus a product of both a particular machine learning algorithm and the data that is used to learn the model.

2.3 Text Data Preprocessing

The application of machine learning algorithms on text requires preprocessing of that text data into a structured format suitable for computations. Natural language processing (NLP) is a field with roots in computational linguistics dealing with the interaction between text

and computers. The main idea of NLP is to parse and modify text data into a format that is processable by a computer, and to then apply algorithms to yield results or insights from raw text that are meaningful to humans (Sarkar 2019).

2.3.1 Vector Space Representation of Text

A collection of documents, such as a set of news articles, are referred to as a corpus in NLP terminology. A way to represent a document in a corpus is by a vector space model, where each dimension of that vector space maps to a unique term in the entire corpus. Each unique term thus represents a feature in the corpus dataset, such as the binary occurrence or the frequency of a term in a document. A document is represented by these features and the entire representation of all documents in the corpus is referred to as a bag-of-words model, as it disregards term sequences, grammar, and order. The entire corpus can then be represented as a document-term matrix, where each column corresponds to a feature of a unique term and each row corresponds to a document. Each element thus represents a feature of a particular term in a particular document, where this feature can be the term count or some other representative value (ibid.). A document-term matrix of a trivial corpus with features as term counts is illustrated in Figure 2.1.

To transform raw text into a bag-of-words representation, it is common to tokenize (i.e. split) text into separate tokens that represent linguistic units. The delimiter on which to split on is usually whitespaces so that each token represents a term. Removal of dots, commas, and other symbols can also be done in this step. The document then consists of a finite sequence of tokens that can be mapped to the vector space of a bag-of-words model (ibid.). Token and term are used interchangeably throughout the thesis, but the former emphasizes a linguistic unit during text data preprocessing and the latter emphasizes an actual word or word sequence.

Corpus

$\mathbf{d_1}$	a topic model analyzes text
$\mathbf{d_2}$	the model finds a hidden pattern
$\mathbf{d_3}$	the pattern is the topic
$\mathbf{d_4}$	the topic is coherent

Tokenized Corpus

$\mathbf{d_1}$	'a', 'topic', 'model', 'analyzes', 'text'
$\mathbf{d_2}$	'the', 'model', 'finds', 'a', 'hidden', 'pattern'
$\mathbf{d_3}$	'the', 'pattern', 'is', 'the', 'topic'
$\mathbf{d_4}$	'the', 'topic', 'is', 'coherent'

Document-Term Matrix

	a	**analyzes**	**coherent**	**finds**	**hidden**	**is**	**model**	**pattern**	**topic**	**text**	**the**
$\mathbf{d_1}$	1	1	0	0	0	0	1	0	1	1	0
$\mathbf{d_2}$	1	0	0	1	1	0	1	1	0	0	1
$\mathbf{d_3}$	0	0	0	0	0	1	0	1	1	0	2
$\mathbf{d_4}$	0	0	1	0	0	1	0	0	1	0	1

Figure 2.1: An illustrative example of a document-term matrix for a trivial corpus

2.3.2 Dimensionality Reduction and Feature Extraction

The bag-of-words representation inherently leads to computational problems as its representation in matrix form is typically sparse and of high dimensionality (Feldman, Sanger, et al. 2007), relating to the general issue of the curse of dimensionality (Murphy 2012). Various NLP techniques exist for dealing with this, relating to reducing the number of features (i.e. unique terms) by removing or merging irrelevant syntactic aspects of these features. A trivial approach is to remove common terms with low significance called stopwords, such as "the", "me", "to" or other irrelevant terms relating to a specific problem.

Two other common techniques are stemming and lemmatization. Stemming refers to removing inflections of terms to their stem algorithmically, so that for example "run", "running" and "runs" all are stemmed to "run", reducing features of the corpus by two. Inflections with varying stems such as "ran" can be problematic, as it is not merged into the "run" feature even though they are of the same verb. Lemmatization is a similar technique that maps inflections of a term to its lemma, meaning that the above "ran" would be mapped to "run" as it is the lemma of "ran". This process is done by lexicographic lookups, heuristics, or trained machine learning models and is thus slower but generally more accurate compared to

stemming (Sarkar 2019). For germanic languages, lemmatization tends to give better results compared to stemming (Haselmayer and Jenny 2014). Stemming can additionally be applied to an already lemmatized term, but if the lemmatization is accurate then this will yield little to no benefits while mutating lemmas to less human-readable stems.

A limitation of the bag-of-words model is its inability to represent idiomatic phrases of sequences of terms. Representing "Bonnier News" as one feature instead of "Bonnier" and "News" separately when those terms occur together increases the expressiveness of the model. This can be achieved by parsing the corpus and statistically evaluate whether sequences of terms are likely to occur in that sequence versus occurring individually. This process is referred to as forming ngrams, where n stands for the number of terms in sequence. This increases the accuracy of the model to represent terms as real entities but can increase the dimensionality of the model (Mikolov et al. 2013). For example, the trigram "Black Lives Matter" is formed from the term sequence "Black", "Lives", and "Matter", but the terms can still exist individually, thus increasing the dimensionality of the corpus by one.

Depending on the application it can be useful to filter out terms belonging to a specific part of speech (POS) class to create a corpus only with features of interest, such as nouns. This requires a trained model to recognize the POS for terms in a text so that they can be filtered out (Sarkar 2019). Depending on the application, common examples of POS classes that are kept are nouns or adjectives. In the context of topic modeling, Martin and Johnson (2015) showed that a nouns only (common and proper nouns) dataset produced the most meaningful topics. They suggest that reducing the articles to nouns only may be advantageous since this improved the semantic coherence of the topics. Another interesting fact was that even when the used text contained all POS classes, topic modeling still favored nouns as the most frequent terms in the topics. Jacobi, Van Atteveldt, and Welbers (2016) similarly found that filtering terms by POS classes tended to yield more interpretable topics when keeping only common nouns and proper nouns, and potentially verbs and adjectives depending on the use case.

2.3.3 Term Frequency-Inverse Document Frequency Representation

The elements (i.e. features) in a bag-of-words document-term matrix representation of a corpus are often expressed as the frequency or count of a term t_i in document d_j. This is referred to as term frequency. Two concerns arise with this representation. If two documents in the corpus have a similar distribution of terms but are of variable length, then the term frequencies of the longer document will be higher and thus have a higher weight. Secondly, terms that occur frequently in all documents in the corpus are less likely to infer meaningful distinctions between documents but will be assigned a substantial weight of the term distribution in the corpus. To address this, the elements in the document-term matrix can be transformed an represented as term frequency-inverse document frequency (tf-idf or tfidf) values. Term

frequency is the frequencies of all terms present in a single document, optionally normalized across that document. Inverse document frequency measures the extent to which a term is present in all documents in the corpus, and gives unusual terms a higher weight (Qaiser and Ali 2018). Mathematically tf-idf is formulated as the element-wise product of a (possibly normalized) term frequency matrix tf and an inverse document frequency vector idf as in Equation 1.

$$tfidf = tf \cdot idf \tag{1}$$

The term frequency $tf(t_i, d_j)$ represents the frequency of term t_i in document d_j. The inverse document frequency for term t_i, $idf(t_i)$ is defined as in Equation 2 where N is the total number of documents in the corpus, n_i is the number of documents in which the term t_i occurs and the added 1 constants are smoothing terms, following notation by Sarkar (2019).

$$idf(t_i) = 1 + \log \frac{N}{n_i + 1} \tag{2}$$

The tf-idf feature value for term t_i in document d_j is given by Equation 3.

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_i) \tag{3}$$

This tf-idf transformation of all elements in the document-term matrix yields a tf-idf matrix. In this matrix, a term in a document is given a higher tf-idf weight proportional to its frequency in a document normalized by the number of terms in the document, and to how unique the term is with respect to other documents in the corpus. A limitation of the tf-idf matrix representation is that it inherently does not account for sequences, term order, and semantic structure of text but it is nonetheless a powerful representation of a corpus (Qaiser and Ali 2018), and usable as input for machine learning algorithms. In the document-term matrix in Figure 2.1, the tf-idf transformation would lead to tf-idf weights instead of counts in the matrix elements.

2.4 Topic Modeling

Topic modeling is a machine learning method that aims to discover the themes that run through a corpus by analyzing the terms in it. This is done by using statistical relationships between the terms in the documents that the corpus consists of. Topic models assume that there exists a user-specified number of K latent topics in a corpus of N documents with a total vocabulary size V. The corpus input is transformed into an $N \times V$ document-term matrix. The key idea is that a document is made up of multiple topics. Hence the aim is to discover a topic distribution over each document, represented by an $N \times K$ matrix, and a

term distribution over each topic, represented by a $K \times V$ matrix. Each topic is a mixture of terms in the fixed vocabulary. A topic can therefore be viewed as a weighted vector or probability distribution over the vocabulary. Similarly, each document is a mixture of topics (L. Liu et al. 2016). This is an accepted assumption in general since a document in a corpus often combines different ideas or themes that permeate the collection as a whole (Blei and Lafferty 2009). In this sense, a topic is essentially a collection of terms with different weights, and a topic is usually presented as the n most heavily weighted terms, where it is common to use n values of 5 or 10. For example, *patient*, *doctor*, *treatment*, *medicine*, *care* is a top-five representation of a topic where these terms have the highest weight, but all other terms in the vocabulary of size V are also present in the topic, but with lesser weight.

The theoretical explanations of the algorithms that are used in this thesis are presented in the following section. To start with, a background of how the different algorithms emerged will be given. For simplicity, the notations of matrices in Table 2.1 are used.

Table 2.1: Notations of matrices used in topic modeling

Matrix	**Dimension**	**Description**
X	$N \times V$	Document-term matrix of the corpus.
W	$N \times K$	Documents as rows, topics as columns. Topic distribution over documents.
H	$K \times V$	Topics as rows, terms as columns. Term distribution over topics.

2.4.1 Background

The origin of topic model algorithms is latent semantic analysis (LSA), also referred to as latent semantic indexing (LSI) (Deerwester et al. 1990). The application of this algorithm on a text corpus requires that the corpus is first transformed into a document-term matrix, here denoted by X. LSA builds on singular value decomposition (SVD) to factorize this matrix X of the corpus into a set of component matrices. These matrices can be reduced to a lower rank and thus be an approximation of X when multiplied (Xu, X. Liu, and Gong 2003; Casalino, Del Buono, and Mencar 2016). One of these component matrices describe basis vectors, or eigen features, for describing X in possibly lower dimensions, while another component matrix represent a mapping of those bases to describe the data samples in X in the original dimensions (Deerwester et al. 1990).

The conceptual idea of LSA, and topic modeling in general, is to factorize the document-term matrix of the corpus into one matrix containing topic-term information (i.e. basis vectors) and another matrix containing document-topic information (i.e. mapping between basis vectors and X), denoted by W and H in this thesis. The topic-term matrix H describes each topic as a weighted vector of length V, where each weight corresponds to the importance of a term in that topic. The document-topic matrix W describes each document as a weighted vector of length

K, where each weight corresponds to the importance of a topic in that document. LSA serves as the basis on which other, more successful topic model algorithms build. Most topic models share the same composition of a topic-term matrix and a document-topic matrix, and the topic model algorithms aim to derive these after some optimality objective. It is noteworthy that for dimensionality reduction applications it is common to work with the transpose of X^T (a term-document matrix), thus requiring transposes W^T and H^T which leads to other matrix operation orderings for factorization. Different notations are used in different literature but this thesis follows notations in Table 2.1. The structure of these matrices is illustrated in Figure 2.2.

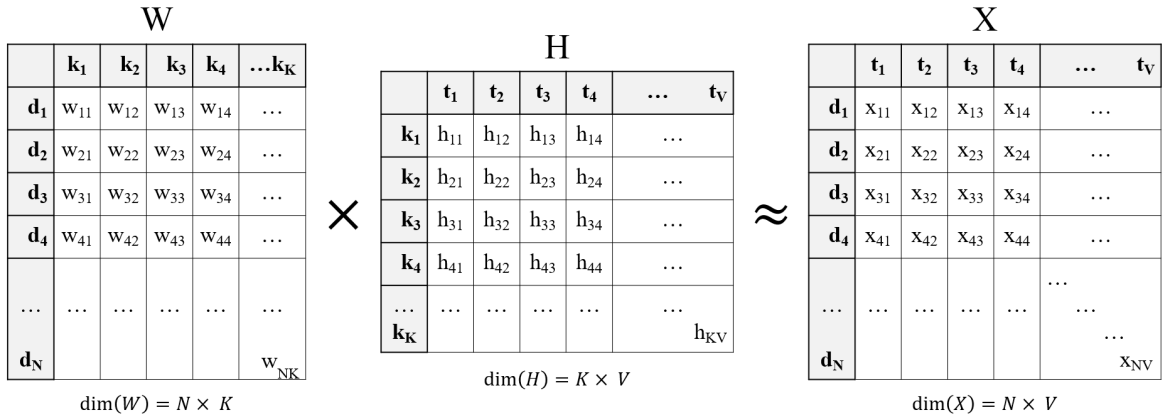

Figure 2.2: Illustrative example of the topic model matrices

The development of topic model algorithms, originating from LSA, have taken two different routes: one probabilistic approach and one approach that builds on linear algebra.

The latter approach resulted in the non-negative matrix factorization by D. D. Lee and H. S. Seung (1999). This algorithm builds heavily on linear algebra and optimization theory by posing non-negative constraints on the factorization of matrices W and H, thus differing from the SVD approach. These constraints create a sparse representation of topics, where only additive operations are allowed to map topics to documents and thus creating a parts-based representation of documents and topics. The NMF topic model algorithm normally takes a tf-idf document-term matrix as input, with tf-idf weights as features in this matrix.

The probabilistic perspective of topic model algorithms starts with probabilistic latent semantic analysis (pLSA). pLSA, also known as pLSI, developed by Hofmann (1999) is a direct extension of LSA that took the development of topic model algorithms in a probabilistic direction. Although Hofmann's work is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents, the per-document topic proportion. The reason is that each document is represented as a list of

numbers, but there is no generative probabilistic model for these numbers. The result is that it is difficult to apply the model to new documents. With the motive to amend this, Blei, Ng, and Jordan (2003) extended this topic model algorithm by introducing latent Dirichlet allocation (LDA). This topic model algorithm is an even more complete probabilistic generative model since the per-document topic proportion is assigned a Dirichlet prior (Steyvers and T. Griffiths 2007). The LDA topic model algorithm normally takes a term frequency document-term matrix as input, with term counts as features in this matrix.

In the following sections, two of the most prominent topic model algorithms to date are discussed, namely non-negative matrix factorization (NMF) and latent Dirichlet allocation (LDA), which are the two algorithms that are implemented and evaluated in this thesis.

2.4.2 Non-Negative Matrix Factorization (NMF)

NMF can be described as an extension to LSA by imposing constraints on the matrix factorization process and thus differing from SVD, as there are notable issues with the SVD representation (D. D. Lee and H. S. Seung 1999). NMF normally takes a tf-idf document-term matrix as input, this representation is explained in section 2.3.

A property of SVD is that the basis vectors will be orthogonal to each other. In achieving this, some elements in the bases are forced to be negative. This causes some interpretative issues when considering the basis vectors as describing features in X. Negative elements in the bases and mappings cause subtractions between columns, leading to a spread distribution of bases in describing a sample in X. Describing data samples as not belonging to a basis (i.e. a topic) by a negative degree due to a negative coefficient in W is also problematic for interpretability (Xu, X. Liu, and Gong 2003).

D. D. Lee and H. S. Seung (1999) described factorizations with negative bases, such as SVD and PCA, to learn a holistic representation of the data, as subtractive combinations are allowed between components. This stands in contrast to a parts-based representation, where parts are summed up to constitute the whole, which better models human conception. Following this reasoning, D. D. Lee and H. S. Seung (ibid.) proposed non-negative matrix factorization, where the data matrix X is factorized into matrices W and H that approximate X with the constraint that W and H only contain non-negative elements. This leads to enhanced interpretability due to non-negative representations of bases and encodings in W and H, and also to an increased sparseness in these matrices as many elements are forced to zero (ibid.). This is desirable as the topic encodings in H will be described by fewer and more distinguishable features (i.e. terms), and the bases that describe assignments of topics to documents in W will also be fewer and more distinguished. The approximation of X by the product WH will be of equal or lower rank K, with $(N + V)K \leq NV$ (Casalino, Del Buono, and Mencar 2016).

The algorithm for deducing W and H from X can be posed as an optimization problem, where the difference D between WH and X is minimized as in Equation 4.

$$min\ D(X; W, H) \qquad \text{s.t.} \quad W \geq 0, H \geq 0 \tag{4}$$

One of the most frequently adopted difference measures is the Frobenius norm, denoted by Equation 5 with its objective function denoted by Equation 6 (D. Seung and L. Lee 2001).

$$||X - WH||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|X_{ij} - (WH)_{ij}|^2} \tag{5}$$

$$min\ ||X - WH||_F \qquad \text{s.t.} \quad W \geq 0, H \geq 0 \tag{6}$$

Another common difference measure is to minimize the Kullback-Leibler divergence, denoted by Equation 7 with its objective function denoted by Equation 8 (ibid.).

$$D(X||WH) = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} \times \log\left(\frac{X_{ij}}{(WH)_{ij} + 1}\right) - X_{ij} + (WH)_{ij}) \tag{7}$$

$$min\ D(X||WH) \qquad \text{s.t.} \quad W \geq 0, H \geq 0 \tag{8}$$

Although the Frobenius norm is most common due to its simplicity and general-purpose use cases, it assumes that the variations in the data are Gaussian due to its least-squares implementation. Chi and Kolda (2012) argues that for sparse count-based data it is more realistic to model variations as Poisson and to instead minimize the Kullback-Leibler divergence for NMF to better capture this. Both objective functions are thus interesting to evaluate, depending on the input data. Regularization by the L1 or the L2 the norm can also be added to the objective function, to further increase the sparseness in the W and H matrices (Hsieh and Dhillon 2011).

The approximation of X by NMF is not straightforward, as the optimization problem is not convex in both W and H due to the non-negative constraints. It is therefore unrealistic to derive a global minimum, but optimization methods can be employed to find a local minimum. These methods take an alternating approach to first update W while H is fixed, and to then update H while W is fixed until the decrease in the objective function between iterations is lower than a threshold ϵ. This is because the subproblems in W and H are separately convex (Gillis 2014). The alternating solver approach can be described as:

1. Generate initial matrices $W^{(0)} \geq 0$ and $H^{(0)} \geq 0$
2. for t = 1, 2, 3, ... do:
   (a) $W^{(t)}$ = update(X, $H^{(t-1)}$, $W^{(t-1)}$)
   (b) $H^{(t)}$ = update(X, $W^{(t)}$, $H^{(t-1)}$)
   (c) if D(X; $W^{(t-1)}$, $H^{(t-1)}$) - D(X; $W^{(t)}$, $H^{(t)}$) $\leq \epsilon$ then stop

(Gillis 2014)

The standard proposed method is the multiplicative update solver (MU) that updates the W and H matrices in turn by matrix multiplication rules and guarantees nonincreasing updates in the objective function (D. Seung and L. Lee 2001). These update rules are described in Equations 9 and 10.

$$W^{(t)} \leftarrow W^{(t-1)} \times \frac{(XH^T)}{(W^{(t-1)}HH^T)} \tag{9}$$

$$H^{(t)} \leftarrow H^{(t-1)} \times \frac{(W^T X)}{(W^T W H^{(t-1)})} \tag{10}$$

Another common NMF solver is the coordinate descent (CD), which selects one coordinate, or a block of coordinates, in the objective function at a time and updates along these coordinate axes. The non-selected coordinates are fixed and the gradient of the hyperplane of the resulting coordinates is calculated which is used with a line search to update the objective function in the selected coordinate directions. Likewise, the W and H matrices are updated one at a time, where the other remains fixed (Hsieh and Dhillon 2011).

There are arguments for the superiority of NMF compared to SVD (and thus LSA) in generating interpretable, separable, and more coherent topics and topic assignments to documents (D. D. Lee and H. S. Seung 1999; Xu, X. Liu, and Gong 2003; Casalino, Del Buono, and Mencar 2016). This is not without drawbacks though. The non-negative constraints make the approximation of X more difficult to achieve and can thus lead to inaccuracies in the resulting components. Finding the exact solution to NMF is NP-hard in general and thus computationally infeasible for practical scenarios. Moreover, NMF is an ill-posed problem, meaning there usually exists several solutions W, H, and $W^{'}$, $H^{'}$ that yields an equivalent approximation of X leading to possibly different outcomes for different runs with the same X input, given different initializations of the origin matrices W and H. Lastly, the user-specified parameter K (the matrix factorization rank and number of topics) is also a nontrivial selection (Gillis 2014).

2.4.3 Latent Dirichlet Allocation (LDA)

Whilst the development of NMF steered towards a linear algebraic approach to topic modeling, latent Dirichlet allocation (LDA) took a probabilistic approach to the same problem. LDA introduced by Blei, Ng, and Jordan (2003), is a generative probabilistic topic model algorithm. LDA is referred to as a latent (hidden) variable model. Latent variable models structure distributions for how observed data interact with hidden random variables, and for LDA this means that a topic structure exists with hidden random variables. When making inferences in a generative model, the goal is to find the best set of latent variables that can explain the observed data.

The input data is represented as a bag-of-words document-term matrix. See section 2.3 for more information about this representation. For LDA, each cell in the document-term matrix represents the term count. In practice, this means that each document in the corpus is represented by a vector where each vector element represents the count of occurrences of a specific term in that document. The output, the topic structure is, as in NMF, defined by the topic distribution over each document W, represented by an $N \times K$ matrix, and the term distribution over each topic H, represented by a $K \times V$ matrix (Blei 2012).

To explain the idea of LDA, some additional notations will be introduced. Suppose d denotes a document, k is a topic and t represents a term. Then, $p(k|d)$ denotes the probability of topic k in document d, and $p(t|k)$ means the probability of term t in topic k. In LDA, both these distributions are assumed to be multinomial distributions. To simplify notations, let φ_k refer to the multinomial distribution over terms for topic k, $p(t|k)$ and let θ_d refer to the multinomial distribution over topics for document d, $p(k|d)$.

Furthermore, the Dirichlet distribution is introduced. The idea in LDA is to place a Dirichlet distribution $Dir(\alpha)$ on the multinomial distribution θ_d, and another Dirichlet distribution $Dir(\eta)$ on the multinomial distribution φ_k. The reason to use these Dirichlet distributions is that this distribution is a conjugate prior for the multinomial. A conjugate prior means that if the prior distribution of the multinomial parameters is a Dirichlet distribution, then the posterior distribution is also a Dirichlet distribution. The benefit of this is that the posterior distribution is easy to compute, in other words, it simplifies the statistical inference (L. Liu et al. 2016). Both α and η are hyperparameters, and good choices of these will depend on the number of topics and vocabulary size. The hyperparameter α can be interpreted as a prior observation count for the number of times a topic is sampled in a document, before having observed any actual terms from that document. The hyperparameter η can be interpreted as the prior observation count on the number of times terms are sampled from a topic before any term from the corpus is observed (T. L. Griffiths and Mark Steyvers 2004).

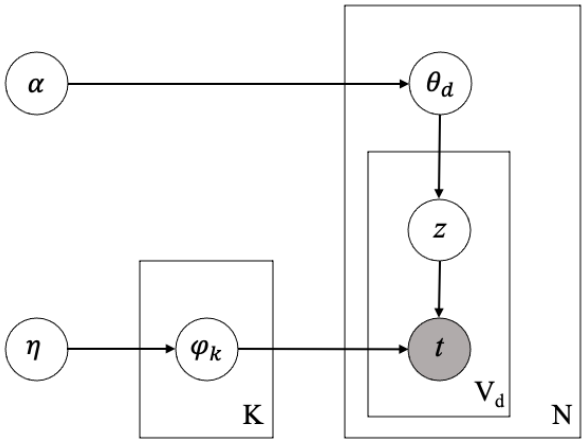

Figure 2.3: LDA graphical model

The generative process in LDA can be illustrated as the directed graphical model in Figure 2.3. In this graphical notation, shaded variables indicate observed variables and unshaded variables indicate unobserved (latent) variables. Since each topic is a mixture of terms in the vocabulary, z corresponds to this term assignment for a specific topic. The distributions φ_k and θ_d, as well as z are the three latent variables that should be inferred. Arrows indicate conditional dependencies between variables, and plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples. For example, the inner plate over z and t illustrates the repeated sampling of topics and terms until V_d terms have been generated for document d. The plate surrounding θ_d illustrates the sampling of a distribution over topics for each document d for a total of N documents. The plate surrounding φ_k illustrates the repeated sampling of term distributions for each topic k until K topics have been generated (T. L. Griffiths and Mark Steyvers 2004).

The generative process can be described as:

1. for each topic $k \in \{1, \ldots, K\}$:
   (a) draw a distribution of terms φ_k
2. for each document $d \in \{1, \ldots, N\}$:
   (a) draw a vector of topic proportions θ_d
   (b) for each term in document d:
      i. draw a topic assignment z
      ii. draw a term t

(Blei and Lafferty 2009).

As mentioned, the distributions φ_k, θ_d, and z are the three latent variables that should be inferred. The generative process defines a joint probability distribution over both the observed and latent variables. The inference is performed by using that joint distribution to compute the conditional distribution of the hidden variables, given the observed variables. This conditional distribution is also called the posterior distribution.

For an LDA algorithm, the joint distribution of a corpus is given by Equation 11 with topics $\varphi_{1:K}$, where each φ_k is a distribution over terms, the per-document topic proportions θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d. The topic assignments for document d are z_d, where $z_{d,m}$ is the topic assignment for the mth term in document d. Finally, the observed terms for document d are t_d, where $t_{d,m}$ is the mth term in document d.

$$p(\varphi_{1:K}, \theta_{1:N}, z_{1:N}, t_{1:N}) = \prod_{d=1}^{N} p(\theta_d) \prod_{k=1}^{K} p(\varphi_k) \prod_{m=1}^{M} p(z_{d,m}|\theta_d) p(t_{d,m}|z_{d,m}, \varphi_{1:K}) \tag{11}$$

The posterior distribution $p(\varphi, \theta, z|t)$ represents the distribution of the topic structure given the observed documents and can be estimated via the joint distribution. Using the notation above, this posterior distribution is given by Equation 12 (Blei 2012).

$$p(\varphi_{1:K}, \theta_{1:N}, z_{1:N}|t_{1:N}) = \frac{p(\varphi_{1:K}, \theta_{1:N}, z_{1:N}, t_{1:N})}{p(t_{1:N})} \tag{12}$$

Computing this posterior is the goal of the algorithm. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure. However, the number of possible topic structures is exponentially large, and in practice, this distribution needs to be approximated. There are a few classic approaches to making inferences in LDA, generally divided into two categories: sampling-based algorithms and optimization-based algorithms (ibid.). In this thesis, two inference methods for LDA are proposed, namely Gibbs sampling as a sampling-based algorithm and online variational Bayes as an optimization-based algorithm.

Gibbs Sampling

Gibbs sampling (Steyvers and T. Griffiths 2007) begins with estimating z given the observed terms t, while marginalizing out φ_k and θ_d, and then approximate φ_k and θ_d using posterior estimates of z. Gibbs sampling is a variant of a Markov chain Monte Carlo (MCMC) algorithm. MCMC is a set of approximate iterative techniques developed to sample values from complex distributions. Gibbs sampling simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others. This is done by constructing a Markov chain, which is a sequence of random variables, each dependent on the previous. In inference for LDA, the Markov chain is defined on the topic term assignments z, and the algorithm samples from the posterior over z by repeatedly

sampling z conditioned on the observed variables. The sampling is done sequentially and proceeds until the sampled values approximate the target distribution. After inferring a posterior distribution of z, this distribution is used to approximate the distributions φ_k and θ_d.

Online Variational Bayes

Online variational Bayes is a deterministic alternative to sampling-based algorithms. Instead of approximating the posterior with samples, this algorithm places several distributions over the latent variables and then finds the distribution that is closest to the posterior with an optimization approach. Online variational Bayes is based on variational inference, which is called variational Bayes (VB). The idea in VB is to optimize the distribution to be close in Kullback-Leibler divergence to the posterior. VB requires a full pass through the entire corpus each iteration, and can therefore be very slow to apply if the corpus consists of a lot of documents. Online variational Bayes was proposed to make this process more effective and is based on online stochastic optimization, which has been shown to produce good parameter estimates dramatically faster than traditional VB on large datasets (Hoffman, Bach, and Blei 2010).

2.5 Evaluation of Topic Models

Evaluating topic models is challenging due to their unsupervised learning process. There is no correct list of topics to compare against for every corpus which could serve as a benchmark or for measuring the error rate. This has led to an increasing interest in the field of measuring the quality of topic models, and a lot of research has been made in creating frameworks to solve this problem, but it remains an open research area (Röder, Both, and Hinneburg 2015).

When using a topic model, the primary concern is the degree to which the learned topics match human judgment, as this is the goal for most use cases. Since evaluations of topic models by humans are extremely time-consuming, the goal is to find measurements that correlate the most with human judgment (Chang et al. 2009). Topic coherence is introduced in the following sections, which is one such measurement.

2.5.1 Topic Coherence Metrics

Topic coherence metrics score a single topic by measuring the degree of semantic similarity between high probability terms in the topic. These metrics are used to distinguish between semantically interpretable topics and topics that are arbitrary artifacts of statistical inference, where the first stated option usually is the most sought after (Stevens et al. 2012).

According to Röder, Both, and Hinneburg (2015) and Mimno et al. (2011) the metrics Cv and UMass, are two metrics that have been shown to match well with human judgments of topic quality. Both measures are based on the same high-level idea, to compute the coherence of a topic as the sum of pairwise scores over the set of topic terms, expressed in equation 13.

These scores can be interpreted as how well the terms support each other according to their similarity in respect to all other terms. The median or the average of the individual topic coherence is usually calculated to measure the topic coherence of the full model. Here T is the set of topic terms and epsilon indicates a smoothing factor.

$$coherence(T) = \sum_{(t_i,t_j)\in T} score(t_i, t_j, \epsilon) \tag{13}$$

The topic terms, $t_1, \ldots, t_n$ are usually the top n terms for each topic. The top terms means the terms with a high probability of describing the topic. The most common way is to use the top ten terms, which captures topic quality with the highest correlations with human judgment (Röder, Both, and Hinneburg 2015).

An important distinction between UMass and Cv is that UMass should be used together with the corpus that was used during the creation of the model, hence calculates this score function with an intrinsic approach. The Cv metric should be used together with an external reference corpus for calculating the score function, which makes Cv an extrinsic metric. Ideally, both techniques (intrinsic and extrinsic) should be used together to have a good picture of the model in terms of coherence. The reason is that they reflect different aspects of interpretability and tend to produce different results since an intrinsic approach captures how well the terms confirm each other in respect to the document that was used when creating the topics, while an extrinsic measure uses a large external corpus and hence capture how well the terms confirm each other in a more general sense (Stevens et al. 2012).

UMass Metric

The UMass metric (Mimno et al. 2011) defines the score to be based on document co-occurrence, specified in equation 14.

$$score(t_i, t_j, \epsilon) = \log \frac{D(t_i, t_j) + \epsilon}{D(t_j)} \tag{14}$$

Here $D(t_i, t_j)$ counts the number of documents containing the terms t_i and t_j, and $D(t_j)$ counts the number of documents containing t_j. As mentioned, these counts are over the original corpus. Each term $t_1, \ldots, t_n$ is compared to the preceding and succeeding term respectively, where the terms are ordered by the probability of describing the topic. The mean of the confirmations measures is then calculated to a single coherence score for each topic, and this is the final UMass coherence score (ibid.).

Cv Metric

Calculation of the Cv metric (Röder, Both, and Hinneburg 2015) is performed as follows: First, the algorithm creates term pairs by taking all the other top n terms for each of the

top n terms in the topic. For example, if $T = \{t_1, t_2, t_3\}$, then a pair is $S_i = \{T^{'} = (t_1), T^* = (t_1, t_2, t_3)\}$. This segmentation measures the extent to which the subset T^* supports, or conversely undermines, the subset $T^{'}$.

Then, for every pair of term subsets, $S_i = \{T^*, T^{'}\}$, calculation of a confirmation measure is performed of how strong the conditioning set of terms T^* supports $T^{'}$, and is based on the similarity of T^* and $T^{'}$ in respect to all terms in T. The similarity between individual terms is calculated via normalized pointwise mutual information (NPMI), as shown in Equation 15. A small constant ϵ is used to account for the logarithm of zero and γ to place more weight on higher NPMI value.

$$\text{NPMI}(t_i, t_j)^\gamma = \left(\frac{\log \frac{p(t_i, t_j) + \epsilon}{p(t_i) \times p(t_j)}}{-\log(p(t_i, t_j) + \epsilon)} \right)^\gamma \tag{15}$$

The joint probability of two terms $p(t_i, t_j)$, is calculated with a boolean sliding window algorithm. This algorithm tries to capture probabilities in respect of the frequencies and distances of terms, and not only the times the term occurs in the documents. This is done by sliding a window over the documents, one term per step. Each step defines a new virtual document by copying the window content, which is used to calculate to compute the term probability. The documents used in this sliding window approach should come from a large external reference set for Cv to be an extrinsic measure.

Finally, the mean of the confirmations measures is calculated to a single coherence score, and this is the final Cv score (Röder, Both, and Hinneburg 2015).

3 Related Work

This chapter highlights previous research of topic modeling and other techniques with a similar purpose to give context to this thesis. First, related methods and content analysis in general is discussed to give a broader perspective of the conceptual task of topic modeling to be able to compare it with other techniques. Next, previous work on topic models applied to news data is outlined to give an indication of how these models can provide value in a news media context and the challenges they face. Finally, an overview is given of how the LDA and NMF algorithms differ in performance, with emphasis on use cases in practical scenarios. The results and conclusions from previous research mentioned in this chapter will be related to the results of this thesis and discussed in chapter 7 to answer the posed research questions.

3.1 Topic Modeling as a Content Analysis Method

Content analysis is a broad scientific method with both automated and manual approaches that attempt to analyze patterns in collections of entities, often large volumes of texts (Neuendorf and Kumar 2015). Topic modeling can be viewed as one automated method for that purpose, but it exists in a field of methods with a similar purpose. For content analysis of text data, Grimmer and Stewart (2013) emphasizes that automated models of language are inherently incorrect and at best an approximation of the vastly more complex nature of language as understood by humans. From their perspective, automated methods should be augmentations of manual analysis but do not replace it. Human-based methods are time and resource-intensive, however, which is why automated methods are highly interesting. In their work, they discuss supervised and unsupervised content analysis techniques and their advantages and limitations for manual analysis.

Supervised methods strength lies in the fact that they offer control in what categories to classify texts into, and does so successfully given that there are significant patterns in the input that can be distinctly mapped to these categories. The drawback is that predefined categories need to be known beforehand, or inferred by human expert knowledge (ibid.). Quinn et al. (2010) relate supervised methods with human coding tasks, as manual labeling has to be done beforehand. Like manual tasks, supervised methods have a high cost at startup and suffer from the issue that human labeling of documents can be inconsistent.

Unsupervised methods shift the burden of determining predefined categories and labeling training data beforehand, to validating the model output afterward. The strengths of these methods are low startup costs, and that they discover patterns in texts that are not prespecified or known beforehand (ibid.). Unsupervised methods reduce the information in large text collections and is a substantial simplification. A mathematical good fit on a preprocessed corpus does not necessarily imply a meaningful result, even though the mathematics can indicate the statistical significance of the result. The usefulness of a model lies in the motivations of the user where the goal is to reveal substantively interesting information, but there is no

guarantee that the method will return conceptually interesting clusters. Evaluating unsupervised models in both quantitative and qualitative aspects is therefore an important part of their success and highly motivated (Grimmer and Stewart 2013).

Different datasets and different research questions determine what model is most suitable for the task, particularly so for text models. Grimmer and Stewart (ibid.) highlight that selecting a single optimal model for one task independently of the dataset characteristics is often misguided. The authors make the point that supervised and unsupervised methods should not be viewed as competitors. The methods are instead most productive when used as complements to each other. The authors note the importance of evaluating any type of model results by human expert opinion, and that these models should be viewed as tools to aid humans in analytical work.

Prominent supervised NLP models such as BERT (Devlin et al. 2018) or ELMo (Peters et al. 2018) have recently shown to excel in various supervised NLP tasks, such as text classification. These methods can be very effective for mapping features to predefined labels, but they still suffer from the limitations of supervised methods mentioned above. Unsupervised methods have also been widely attempted for these tasks. Grimmer and King (2011) applied hundreds of unique clustering techniques on text data with successful outcomes but notes that a mathematically optimal inter-cluster and intra-cluster distances do not necessarily imply meaningful clusters. Xu, X. Liu, and Gong (2003) explain that standard clustering techniques often make erroneous assumptions about document distributions when clustering texts, and turn to topic modeling as a better approximation for creating meaningful human clusters. Similarly, a multitude of other research indicates that topic modeling is a state-of-the-art method for automated content analysis (Quinn et al. 2010; Grimmer and Stewart 2013; Stevens et al. 2012).

Validating unsupervised methods is difficult and topic modeling is no exception. Many practitioners in this field highlight the need to use qualitative human measures, or sound quantitative measures to approximate human judgment (Chang et al. 2009; Lau, Newman, and Baldwin 2014; Röder, Both, and Hinneburg 2015). Quantitative metrics can be good proxies for human judgment but they are not perfect. In practice, topic models on large corpora will have to deal with documents that do not fit into any topic category but that are rare enough not to create separate clusters. How to deal with these outliers is an open question and dependent on the use case (Quinn et al. 2010). The conclusions from these works show that topic modeling can be used for content analysis, and potentially as an unsupervised categorization framework for news articles, but that there are clear concerns with model selection and model validation.

3.2 Topic Modeling of News Articles

Topic modeling has previously been applied on news article texts in numerous studies. The aim is often to make the topics represent semantic concepts and to categorize articles into them. This is not always the end result however as these algorithms build topics and relationships based on statistical term co-occurrences. Jacobi, Van Atteveldt, and Welbers (2016) emphasizes this issue and suggests that topics could potentially be formed from other patterns such as writing styles, specific events, or framed concepts. They highlight that how a topic is interpreted is in essence an empirical question with no objectively true answer, thus leaving a topic model's usefulness highly related to the context in which it is used.

The K parameter, specifying how many topics to extract from a corpus, is a key adjustable hyperparameter of a topic model in its usefulness for a particular use case. Quantitative metrics can be optimized to find an optimal K, but a mathematical best number of topics does not necessarily imply the most meaningful model (Chang et al. 2009). Metrics such as coherence are known to best simulate human judgment in this matter (Röder, Both, and Hinneburg 2015), but ultimately the K number of topics to find is highly related to the use case context of the topic model. The goal is to describe the data in fewer dimensions by the K topics but to do so while losing as little information as possible (Jacobi, Van Atteveldt, and Welbers 2016).

Stevens et al. (2012) attempt to answer the question of how many topics to extract by doing experiments over a topic range from 1 up to 500 on a large corpus of over 50 000 New York Times (NYT) articles from nine different NYT sections. LDA, NMF, and LSA algorithms were evaluated, and the study indirectly concludes that all these can produce valuable, interpretable topics in a news article context. The authors evaluate by coherence metrics, term similarity by human judgment, and classification potential on unseen data. For extrinsic coherence metrics, they find that the optimal metrics are gained for K values below 100. For human judgment on term similarity, the similarity increases drastically up to when K is around 100 and then to flatten out or plateau. For classification, the accuracy increases sharply until K is about 50 and then plateau beyond that. The different algorithms used show similar trends in all these aspects, with slight deviations from each other (ibid.). The study indicates that overall, K is best suited between 50 and 100 topics to yield the most interpretable results to a human user.

Jacobi, Van Atteveldt, and Welbers (2016) find the best-suited number of topics to be between 25 to 50 topics when modeling historical newspaper articles about nuclear power plants. The topics in their study mostly resembled particular events with some relation to each other or more general issues over a longer time period. Interesting patterns do surface in their topic results but the authors note a lack of ability to control or tune the semantics of the topics (i.e. if the topic should represent concepts, events, or perspectives for example). They advocate topic modeling as a powerful tool for text analysis but acknowledge that the topics can be

quite random in the concepts they represent (Jacobi, Van Atteveldt, and Welbers 2016). This indicates that topic modeling is a highly explorative method and that the value it provides depends heavily on the research questions the method is supposed to answer.

This is showcased by studies where topic modeling has been used to answer more specific research questions by formulating new or assessing existing hypotheses about a text collection. Chandelier et al. (2018) studied the perspective on wolf recolonization in France in the media over a certain time period by applying topic modeling on French newspapers. By collecting French news articles from that period that to some extent are about wolves, they studied the main wolf-related themes in this corpus. They aimed to understand what concepts or issues are discussed together with the concept 'wolf recolonization in France', and which of these concepts are most dominant. The authors conclude that topic models gave meaningful results and that these models can be effective in content analysis tasks. They assess the limitation that it is difficult to verify that sentiments found in topics relate to wolves and not to farmers that are having issues with wolf-related problems, for example. The authors encourage the use of topic models in content analysis tasks of media coverage, in particular on environmental concerns (ibid.).

In a similar study, Q. Liu et al. (2019) used topic modeling to assess media coverage of third-hand smoke (THS) and what concepts are often related to it in American and Chinese news articles. Similarly, they filtered out articles about THS, segmented them into an American subset and a Chinese subset, and applied topic modeling to each set to see what other concepts, such as cancer, are covered by each media brand and to what extent. Their findings show that American media covers THS and concerns about it to a greater extent compared to Chinese media in this data, revealing significant biases between these contexts. The authors conclude that future research on topic models on news media is strongly warranted (ibid.).

These works show that there are clear motivations for using topic modeling to explore news articles, but that the results can be uncontrollable and that topics can allude to different types of concepts depending on the data input. These are valuable insights for discussing topic model algorithm generalizability between different datasets, as well as for content analysis and categorization framework potential.

3.3 LDA versus NMF

The LDA and NMF algorithms have both been widely used but there is no optimal algorithm in general topic modeling applications. Stevens et al. (2012) present an extensive comparison of LDA, NMF, and LSA evaluated on a large corpus of English news articles. For coherence, they found that LDA slightly exceeded NMF on average. NMF achieved higher coherence scores for its best topics but had a high variance in coherence scores, especially so for K values greater than 100 where NMF learned many low-quality topics. For the term similarity task, LDA exceeded NMF by a slight margin. Additionally, they measured the classification

potential of NMF, LDA, and LSA models on unseen documents and the correlation between their classification accuracy and coherence scores. NMF exceeded LDA in this task by a small margin and had a higher correlation between coherent topics and classification potential. The algorithms showcase different strengths and weaknesses, where even LSA surpassed the other two in certain tasks. NMF performed better in classification but LDA tended to learn more coherent topics. The authors conclude that when the topics should be presented to a human end-user, LDA is more likely to give good results due to its flexibility and stable coherence (Stevens et al. 2012).

By the same reasoning, M'sik and Casablanca (2020) conclude that LDA is a more relevant alternative than NMF for human end users when modeling on a corpus of 13 000 covid19 articles with a small number ($K < 11$) of topics. They show that LDA achieves higher coherence scores by the Cv metric compared to NMF, and they also state that LDA has more meaningful terms in its topics compared to NMF. Suri and Roy (2017) applied topic models to large Twitter text streams for event detection with successful results for both LDA and NMF. Similarly, they conclude that LDA yields more coherent and semantically interpretable topics compared to NMF, but that NMF is significantly faster.

Chen et al. (2019) compared LDA and NMF on five large datasets of short texts of less than 14 terms, mostly on news headlines. They conducted quantitative coherence evaluation of NMF and LDA models and evaluated the topics and topic-article assignment by expert human judgment. In a prelude to their study, they review that probabilistic topic models are generally more popular due to adjustable priors to better mirror specific distributions of topics and terms over documents. LDA and similar methods are assumed to perform best on the condition that the volume of input data is large enough. On short texts, however, this assumption is strongly challenged by their study as the results indicated that NMF is superior compared to LDA both on coherence score and expert human judgment. The authors suggest that since short texts result in a sparse corpus representation, it inherently lacks sufficient information of term co-occurrences for probabilistic models like LDA to be effective. As NMF uses a tf-idf weighting, it contains more prior information compared to LDA. This is especially useful for sparsely represented corpora and leads to NMF being able to produce more high-quality topics and more stable results in general in such scenarios. The NMF algorithm is also more reproducible as it uses deterministic optimization algorithms compared to the stochastic inference algorithms in LDA. Chen et al. (ibid.) also show that the use of external knowledge should be exploited if possible for these algorithms to perform better, especially on sparsely represented corpora such as short texts. They conclude that further comparative research between NMF and LDA on longer texts is strongly warranted.

It should be noted that NMF is generally considered to be significantly faster and more time-efficient than LDA in most settings, especially if LDA is performing inference with Gibbs sampling, which is a slow process (Suri and Roy 2017). NMF has more well-established

strategies for increased efficiency and scalability, where variations of the algorithm have been developed for this purpose (Gillis 2014; Du et al. 2017).

In summary, LDA tends to produce more coherent and human interpretable topics compared to NMF on large corpora with normal-long texts. NMF tends to yield better coherence on sparse corpora, has superior reproducibility, has better classification potential, and is more computationally efficient. There are thus compelling arguments for both algorithms, depending on the use case. These arguments serve as grounds for discussing the best-suited topic modeling algorithm to use in the context of this thesis and will be related to the findings discussed in chapter 7.

4 Data

In this chapter information about the data is provided. The first section gives a general outline of news article text data, followed by a description of specific characteristics for the Swedish language and implications of using Swedish in natural language processing techniques. Then the data used to learn the models in this thesis is presented in more detail, and finally a presentation of the data used as a reference corpus for one of the metrics in the quantitative evaluation of the topic models.

4.1 News Article Text Data

The choice of learning material has a profound impact on how well any kind of natural language model comprehends the language. In this thesis, the models were learned with news article text data, which have some notable features. A corpus that consists of news articles may lead to a variety of problems. First of all, the corpus may be biased, for example, if the media corporation has a clear ideological standpoint, or if the articles are bound to a specific theme, for example, a motor magazine (Lindsey et al. 2007).

Another limitation of text data is the limits of language for data representation. Vast amounts of meaningful information exist in texts of natural languages for a human reader and there is a clear structure in such texts from a linguistic perspective. From a data science perspective, however, texts of natural language are considered highly unstructured as all possible permutations of raw text can not be specified in a predefined schema (Feldman, Sanger, et al. 2007). A consequence is problems when combining the language of journalism with the precision of computation. They are both mechanisms for communicating semantic information, in the form of words, topics, or facts, but do not communicate perfectly with each other. The result is the importance of recognizing the limits of data with respect to language, and the limits of language with respect to data (Caswell 2019).

4.2 Swedish Text Data

Since Bonnier News (BN) is a Swedish media concern, the articles from the brands of BN are written in Swedish. Traditionally, English has been the main language for developing natural language processing techniques, with a result that tools, such as stemming and lemmatization tools, are usually developed for English. It is therefore important to have in mind the characteristics of Swedish, and in which ways Swedish differs from English. Hedlund, Pirkola, and Järvelin (2001) identify five features of the Swedish language that are likely to affect the usage of natural language processing techniques. (1) A fairly rich morphology (meaning structure and content of word forms), (2) gender features, (3) high frequency of compound and derivative word forms, (4) common noun phrases are less frequent in Swedish compared to English, and (5) a high frequency of homographic word form (meaning words that shares the same written form as another word but has a different meaning). Of these, the two char-

acteristics (1) and (3), have been identified to matter most for the natural language processing techniques used in this thesis, and are discussed below.

Compared to English, the Swedish language is more complex regarding the inflectional and derivational morphology. Just to take a few examples, nouns can be divided into five declination types according to the plural suffixes they take, i.e., -or, -ar, -er(r), -n, or no suffix. Genitive forms are formed by the suffix -s. Nouns have several inflectional forms and even the stems change, which means that simple indexing and matching methods are unreasonable to use for the Swedish language, hence stemming and lemmatization is problematic (Hedlund, Pirkola, and Järvelin 2001).

Swedish is characterized by a high frequency of compound words, for example *maskininlärning* (machine learning), which has big implications for using Swedish in natural language processing techniques. Many nominal compounds are lexicalized, hence part of the Swedish language, but since the words are embedded, the component words need to be decomposed to be identified, for example, search keys. Also, a typical feature of Swedish is the use of fogemorphemes in compound word-formation, for example, -s (rätt**s**fall, legal case) or -e (flick**e**barn, female child). There are cases where the word preceding the fogemorpheme is a stem, sometimes a base form, with a result that to be able to decompose these compound words, the fogemorphemes needs to be handled correctly (ibid.).

4.3 Datasets

The data used to learn the models in this thesis are collected from Bonnier News's (BN) data warehouse, which was accessed via Google BigQuery. The data warehouse holds the text for the articles published by brands from BN, and a subset of these was used for creating the corpora in this thesis. A criterion that the character length of the text in the article needed to be greater than 300 was chosen. The reason was to only select actual articles and not news flashes, captions, etc., that also exist in the articles table in the data warehouse.

To fulfill the research aim of this thesis to evaluate different topic models and how well the models generalize to article datasets of multiple news brands, a collection of datasets was extracted from the data warehouse to be able to learn different models on the same data for a more reliable model evaluation process. The articles in these datasets are all published between 2019-01-01 and 2020-02-28. The reason not to choose the most up to date news articles was because of the covid19 pandemic, which had a big influence on articles written after February 2020, and a decision was made to not use news articles extremely influenced by a pandemic, leading to a majority of the data about the same topic as input when learning the models.

Three datasets were collected, all in the same size of 50 000 articles but each of a different type, see Table 4.1. The first type consisted of news articles only from Dagens Nyheter

(DN), referred to as the single brand dataset. The articles were randomly sampled from the previously mentioned time period. This dataset was used to evaluate which topic model algorithm and data preparation yielded the most meaningful topics and categorizations of Swedish news articles within one news brand. The second type was a data set that consisted of news articles from DN, Dagens Industri (Di), and Helsingborgs Dagblad (HD), resulting in data from one large morning paper, one industry magazine, and one local newspaper. The articles were randomly sampled within the three brands, with the dataset as a whole consisting of one-third of articles from each brand. The last type consisted of news articles from 21 of BN's brands. There are today more brands within Bonnier News, but these 21 brands were the ones that had their data stored in the data warehouse during the time period chosen. A list of all the brands included can be found in Appendix A. To get a fair BN representation, this dataset was chosen to consist of one-third of articles from large morning papers, one third from industry magazines, and one third from local newspapers. The reason to use the second and third data set type was to evaluate how well topic model algorithms generalized to article datasets of multiple news brands. The second and third type was referred to as multi-brand datasets.

Among the three types, there might exist overlapping articles since for example articles from DN exist in all of the three datasets. Although, this was not seen as a drawback.

Table 4.1: Datasets used in the evaluation of topic model algorithms in this thesis

Size (number of articles)	**Dataset type**
50000	DN
50000	DN/Di/HD
50000	BN

In addition to the datasets presented in Table 4.1, smaller datasets of various sizes, with only articles from DN, were used for testing and as input in the algorithm selection phase, presented in section 5.3. The reason to use these test datasets was that having large datasets with 50 000 articles when testing different algorithm versions would have been time-consuming and hence an ineffective way of working.

4.4 Extrinsic Data

In section 2.5.1 an explanation of coherence was given, and the use of an external dataset, also known as extrinsic data, was motivated. To summarize this, the extrinsic data was used to calculate the coherence for the Cv metric, to capture how well the top terms in a topic confirm each other in a general sense, meaning not in the context of the data used when

creating the topics. Since the aim was to create good topics in the context of Swedish news articles, the extrinsic data was chosen as 20 000 news articles from 21 of BN's brands, the same brands as for the BN dataset. The time period was chosen as the year 2018. This data was extrinsic since there were no overlapping articles with the articles used for creating the model since the time periods for the article's publication date differ. The extrinsic data did originate from the same brands as for the BN set which can be considered as a bias. However, since the extrinsic data is used as a reference set for representing the Swedish language in a news context and not as a strict validation dataset, this was not considered to be a major limitation. Ideally, the extrinsic dataset should consist of articles from other Swedish news brands, but no such dataset was available for this thesis. The extrinsic dataset was considered large enough to capture the Swedish language in a general sense. The criteria for the character length of the text in the articles was raised to 600 because that gave longer articles with richer content.

5 Method

This chapter presents the method and its underlying methodology for building and evaluating meaningful topic models. First, the premises and goal of the method are stated. Then, an overview of the different parts in a proposed data-model-topics pipeline is presented and explained, followed by detailed outlines of each part.

5.1 Overview

The purpose of this method is to structure a process for learning and evaluating topic models that derive a number of topics so that each document in a specified corpus can be meaningfully described by one, or a few, of those topics. The terms in each topic should be meaningfully coherent and have little overlap with other topics. Moreover, this model should generalize sufficiently well when given corpora of similar content, in other words, Swedish news articles. In this case, meaningful means that the results are interpretable and useful to analysts and domain experts in the field of Swedish news media. Their qualitative judgment is therefore the most important criterion for the validity and usability of the results. However, human involvement in the early stages of choosing a suitable topic modeling algorithm is ineffective, so quantitative measures need to be employed in this stage as an approximation. By optimizing suitable quantitative metrics of the model outputs, the vast space of NMF and LDA algorithms was reduced to a workable set. The output from the models learned by these algorithms was then evaluated by human judgment as a final stage to find the most meaningful model.

As mentioned in chapter 2, there are several different algorithms for topic modeling. In this thesis, two of them were evaluated, LDA and NMF. See section 2.4 for detailed explanations of the theories behind these topic modeling algorithms. Several hyperparameters can be tuned for both of these models. The motivation for evaluating both algorithms with different hyperparameter combinations is that there is no clear optimal algorithm for this task judging from previous studies presented in section 3.3, therefore, it is important to try out different algorithms to find the one that learns the best model to use for Swedish news article text.

A data-model-topics pipeline was used to transform the raw article text into topic models. An illustration of the pipeline is presented in Figure 5.1.

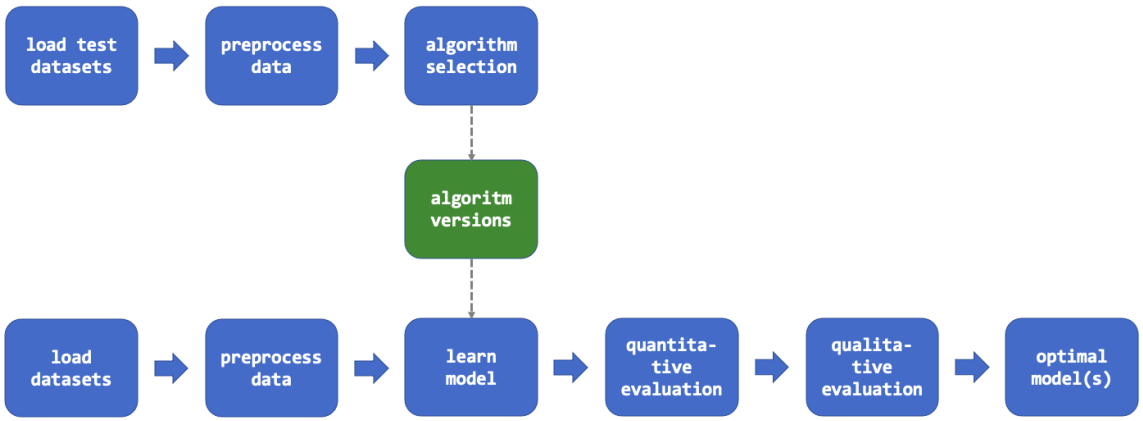

Figure 5.1: An overview of the data-model-topics pipeline, illustrating the method for building and evaluating meaningful topic models

The first phase in the pipeline was to load data, which in the first case was test datasets of various sizes from 1000 to 10 000 articles, all from DN. The news article text data was then preprocessed by cleaning and formatting it suitably for topic modeling. An important choice was which part of speech (POS) classes to use. This choice was not trivial, and a survey was used as a basis for this decision. More information on POS classes can be found in section 2.3.2, and the survey is explained in more detail in the upcoming section.

Next, suitable versions of the NMF and LDA algorithms needed to be found. These versions were found through an explorative algorithm selection phase by manual testing and hyperparameter searches until a few NMF and LDA versions remained. The different versions differed in the way they were implemented, for example performing inference, or in hyperparameter combination. They were chosen based on results from previous related work and manual inspection of the resulting topics.

These versions were then evaluated quantitatively in the next phase. First, new datasets presented in Table 4.1 was loaded and preprocessed. Then, the algorithm versions were used to learn different models to be evaluated over a preprocessed dataset of the single-brand DN dataset type. The model outputs were measured by quantitative metrics to find quantitatively optimal algorithms. The goal was to filter out one LDA and one NMF algorithm. Second, these two algorithms were used to learn models over preprocessed datasets of different types, and the goal was to evaluate how well the two different algorithms generalized when learned with news article text data from multiple news brands.

These two models with corresponding topics were then qualitatively evaluated by human observation and analysis. The goal was to find one optimal topic model that yielded the most meaningful topics to Swedish domain experts, and that generalized well when relearned

with data from multiple news brands. This qualitative evaluation served as a basis for the subsequent discussion and conclusion of the general viability of topic models for describing and categorizing Swedish news articles.

The tools and frameworks used for each part in this pipeline are summarized in Table 5.1.

Table 5.1: Libraries and frameworks used for each part in the pipeline for building and evaluating meaningful topic models

Domain	**Part**	**Library/Framework**
Data Preprocessing	Data Loading	BigQuery
	Data Cleaning and Tokenization	Efselab
	POS/NER tagging	Efselab
	Ngram	Gensim
	Vector Space Representation	scikit-learn
Modeling	NMF	scikit-learn
	LDA	Gensim, MALLET
Quantitative Evaluation	Topic Coherence	Gensim

5.2 Data Preparation

5.2.1 Survey - which Part of Speech Classes is Best to Use in Topics?

In section 2.3.2, part of speech (POS) classes were introduced, and previous studies highlighted the importance of filtering out terms based on their POS class. A survey was conducted with the purpose of finding which POS classes the domain experts at Bonnier News preferred in a topic. The results of the survey were used with previous studies to make a well-motivated decision of which POS classes to keep. Domain experts refer to journalists, editors, product owners, or data analysts that work in the field of Swedish news media.

The survey was designed to present three news articles, and for each article, the domain experts answered two questions. The first question was to write five keywords that they believed described the theme of the article in the best way. In the second question, the domain experts were presented with two different versions of topics that the article was assigned to, one version with a topic from a model that was learned with a corpus that included common nouns and proper nouns, and the other version a topic from a model that was learned with a corpus that included only common nouns. A topic was presented by its top five terms. The respondents were then asked to choose which version they believed described the theme of

the article in the best way. The motivation of comparing the two options proper nouns and common nouns with only common nouns, and not all other POS classes, was that these two options were, by manual inspection and according to previous studies, the most appropriate to use in topic modeling with news article text data.

In the second question, the domain experts were given the option not to choose between the two versions, with an "I do not want to answer this question"-option. This wording was chosen prior to an "I do not know"-option to compel the respondents to make a choice. Since the respondents were asked to choose the best of two versions, an "I do not know"-option was found to be a too easy way to go if the respondents had a hard time deciding. At the end of the survey, the domain experts were given the option to provide comments. The survey was anonymous, but the domain expert had to fill out their role. The surveys were created in Google Forms.

There was a notable problem with this approach, since the answers may depend heavily on what type of articles chosen. The aim was to choose the three articles to minimize the risk of the articles influencing the type of answers too much. Nevertheless, this was a major drawback by only choosing three articles, but seemed like a reasonable amount to present in a survey.

The survey aimed to get an insight into which POS classes that domain experts preferred, partly when they specified keywords they believed described the topic of the article in the best way, partly when they choose between the two options of topics created from corpora from common nouns or common nouns and proper nouns. This information served as a basis for which POS classes to keep in the preprocessing phase introduced in the next section.

5.2.2 Data Preprocessing

The news article text data was loaded from the data warehouse and then processed through the preprocessing phase of the pipeline. This phase is illustrated in Figure 5.2.

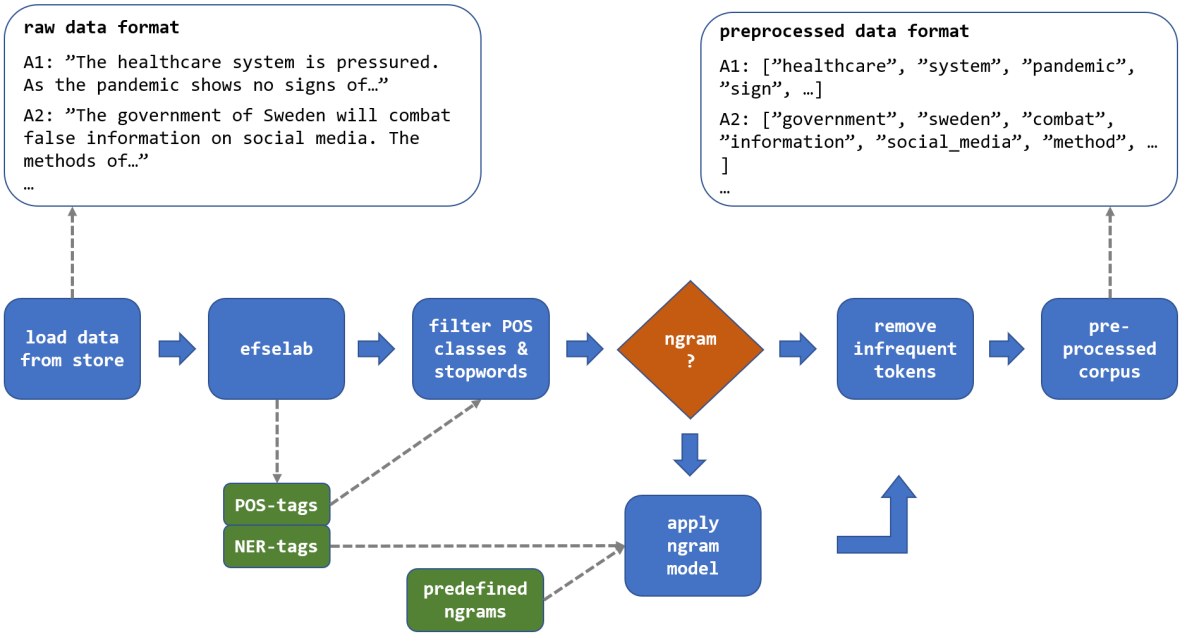

Figure 5.2: Overview of the preprocessing

First, the article texts were processed into a collection of text strings, where each string represented a full article. Each article string was then tokenized and lemmatized, where tokens were cleaned from numbers, punctuations, and other symbols. See Section 2.3.2 for detailed explanations of these techniques. To clarify, a token is a linguistic unit that in this case represents a term. As the text data was in Swedish, these processes were not as straightforward as if they were in English. In section 4.2 the complexity of the Swedish language was explored. This complexity was the first reason for difficulties in lemmatization and other NLP tasks in this project, as the language rules for these processes can be more complex for Swedish. The second reason was that there were few standardized, well-established libraries for these tasks in Swedish.

The main tool used for preprocessing was Efselab (Östling 2020), which implements a tokenizer, a lemmatizer, a part of speech (POS) tagger, and a named entity recognition (NER) tagger for Swedish and other languages. It is partly created by and based on research by Östling (2018). He showed that shallow perceptrons with predefined feature templates for a language can achieve state-of-the-art performance in accuracy on sequence labeling tasks, such as lemmatization, POS, and NER tagging. The gains from using a perceptron based approach, in contrast to common state-of-the-art neural networks for sequence labeling, is that the perceptron approach is significantly faster without necessarily sacrificing accuracy. Other text processing tools, such as UDPipe (Straka 2018) were implemented and evaluated but were either too inaccurate or too slow in processing speed to be feasible to use in this project. Efselab is an accurate and efficient solution and was used for tokenization and lemmatization for the preprocessing part of the pipeline. The output from Efselab is tokenized articles, with

accompanying POS and NER tags for each token.

This output was then cleaned by filtering out stopwords and certain POS classes. The POS classes kept are only common nouns and proper nouns. The reasons to keep these two are motivated by the results presented in section 6.1. The removed POS classes and stopwords are accounted for in Appendix B. Each article was then represented as a list of cleaned, lemmatized tokens.

As described in section 2.3.2, a common approach to create a more descriptive corpus is to form ngrams over frequent sequences of terms. This was explored as an optional add-on to the experiments. The most common ngram approach, as explained in 2.3.2, is to find statistically significant ngrams in the input corpus ad hoc. However, unwanted ngrams tended to be formed when evaluating this approach, such as common titles and names like *president obama*, while other important ngram combinations such as *black lives matters* were missed. This was considered an issue since prefixing the term *obama* with *president* does not yield much meaning and creates two different term features in the document-term matrix, one for single occurances of *obama* and one for occurances of *president obama*, both with the same meaning. Instead, a predefined, more controlled ngram approach was used. A merged dataset of 50 000 articles was analyzed by a module in the library Gensim (Řehůřek and Sojka 2010) to find common ngrams to be used as predefined ngrams. Additionally, the NER tags for each article were used to find sequences of named entities and conditionally merge them if they constituted names with first and last names following each other, or similarly for named entities with double terms, such as *european union*. For each corpus in the pipeline, there was thus an option to apply an ngram model to merge terms by these predefined ngram rules or if two sequential terms represented a named entity.

The corpus was then cleaned of all terms that occur only in a single document. Note that this is not illustrated in the example in Figure 5.2. Terms that occurred less than two times would not be part of any meaningful topic due to their low occurrence, and reducing the corpus vocabulary could improve the performance of the model as well as significantly speed up the computations in the algorithms.

5.3 Algorithm Selection

In the algorithm selection phase, preprocessed test datasets were used as input to the NMF and LDA algorithms to learn topic models. Both algorithms have hyperparameters and inference methods (LDA) or solver methods (NMF) that greatly influence the resulting model. At the outset, suitable hyperparameters and inference or solver methods for the algorithms were unknown. The outcome of the algorithm selection phase was an understanding of how the hyperparameters and inference or solver methods affected model performance, as well as a small set of selected versions to be formally evaluated in the quantitative and qualitative evaluation phases of the pipeline.

The purpose of the algorithm selection phase was to find a set of algorithms of NMF and LDA respectively to be evaluated in the quantitative evaluation phase. These versions were selected in an explorative algorithm selection phase based on results from previous studies and manual inspection of the resulting topics. Input to this phase was datasets of various sizes from 1000 to 10 000 articles, all from DN. Suitable versions were found by a trial-and-error analysis of different implementations and hyperparameter combinations in an iterative process. An automated hyperparameter search was not conducted, due to the reason that this would have been extremely time-consuming. The combinations of hyperparameters were exponentially large and within the scope of this thesis, it was found unreasonable to search through every combination. Instead, a few hyperparameters were considered as the most important, and a search to optimize these were conducted. The resulting versions from this phase were chosen due to differences in inference or solver methods, objective function, regularization, and other hyperparameter combinations to capture varying aspects of LDA and NMF algorithms.

A detailed description of implementations and hyperparameters can be found in section 5.3.1 for NMF and 5.3.2 for LDA, and a discussion of how these affect the model output is given in 6.2.1 and 6.2.2 respectively. In practice, the implementation of the algorithms was done by using two different Python libraries, scikit-Learn (Pedregosa et al. 2011) for NMF and Gensim (Řehůřek and Sojka 2010) for LDA. These libraries were chosen because they provide the best configuration possibilities for the respective algorithms in the scope of this thesis.

5.3.1 NMF Algorithm Selection

The NMF implementation was done by scikit-learn's NMF functionality. The corpus in the list of lists of tokens format was first transformed into a tf-idf matrix by the scikit-learn TfidfVectorizer. See section 2.3.1 for a detailed explanation of this matrix. The relevant input hyperparameters to the NMF implementation are presented in Table 5.2. See the scikit-learn documentation for a more detailed description and the full set of hyperparameters. The output of this method is a W and an H matrix which trivially can be transformed into topics as lists of terms, and topic-article assignments as a vector of weights corresponding to topic relatedness to an article.

Table 5.2: The most relevant NMF hyperparameters

Hyperparameter	**Description**
n_components (K)	Number of topics to be extracted *integer, $k > 0$*
beta_loss	Distance measure to use in the objective function *['frobenius', 'kullback-leibler']*
solver	Optimization method *['cd', 'mu']*
max_iter	Max iterations of algorithm before forced stopping pre convergence *integer, $x > 0$*
init	Initializations for W and H matrices *['random', 'nndsvd', 'nndsvda']*
alpha	Multiplier factor for the regularization term *float, $x \geq 0$*
l1_ratio	Type of regularization. 0 equals only L2 regularization, 1 equals only L1 regularization and in between is a mix of both *float,* $0 \leq x \leq 1$

5.3.2 LDA Algorithm Selection

The LDA implementation was made with the library Gensim, partly with the Gensim built-in algorithm, partly with the Java-based package MALLET, where Gensim provides a Python wrapper. The reason to choose these different implementations was due to how they differ in inference technique. There is no standard best solution for making inference since it heavily depends on the dataset. Since the inference method most suitable for Swedish news article data was unknown, both implementations were found reasonable to evaluate. For both implementations, the output of the models is in the same format as in NMF.

The Gensim algorithm was implemented with the Gensim library, which is a Python library for topic modeling, document indexing, and similarity retrieval with large corpora. The algorithm utilizes the inference technique online variational Bayes. The relevant input hyperparameters to the Gensim implementation of LDA are presented in Table 5.3. See the Gensim library documentation for a more detailed description and the full set of hyperparameters.

Table 5.3: The most relevant LDA Gensim hyperparameters

Hyperparameter	**Description**
num_topics (K)	Number of topics to be extracted *integer, k > 0*
passes	Number of passes through the corpus during training *integer, x > 0*
iterations	Maximum number of iterations through the corpus when inferring the topic distribution of a corpus *integer, x > 0*
alpha	Dirichlet hyperparameter: Document-Topic Density *['auto', 'asymmetric', float]*
eta	Dirichlet hyperparameter: Term-Topic Density *['auto', np.array, float]*

The MALLET implementation was accessed by a Python wrapper for the Gensim library. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text (McCallum 2002). The MALLET implementation of LDA utilizes the Gibbs sampling inference technique. The relevant MALLET LDA input hyperparameters are presented in Table 5.4. See the Gensim documentation for a more detailed description and the full set of hyperparameters.

Table 5.4: The most relevant LDA MALLET hyperparameters

Hyperparameter	**Description**
num_topics (K)	Number of topics to be extracted *integer, k > 0*
iterations	Maximum number of iterations through the corpus when inferring the topic distribution of a corpus *integer, x > 0*
alpha	Dirichlet hyperparameter: Document-Topic Density *integer, x > 0*

5.4 Quantitative Evaluation

The result from the algorithm selection phase was a set of NMF and LDA algorithm versions that differed in hyperparameters and inference or solver methods. This set of algorithms were applied on new datasets presented in Table 4.1 and the learned models were then formally

evaluated in the quantitative evaluation phase of the pipeline. The NMF and LDA algorithms yield similar output formats, where each topic is represented as a list of terms, and each document is represented by a weighted vector of how each topic describes that document. This makes NMF and LDA quantitively comparable. The quantitative evaluation consisted of measuring three distinct metrics over a model's topic and topic-article assignment output, which gave an approximative indication of the quality of that output. The quantitative metrics used were the topic coherence metrics Cv and UMass described in section 2.5.1, and a custom relative sparseness metric (RS) that measures how well the most probable topic describes an article relative to how well all topics describe that article. The implementation of the Cv and UMass metric, as well as more details of the RS metric and its implementation, is presented in 5.4.1 and 5.4.2 respectively.

Both NMF and LDA algorithms share the important hyperparameter K, which specifies the number of topics to find. Higher K values generally give more granular topics and lower K values give broader topics. As there were use cases for both granular and general topics in the context of this thesis, it was interesting to measure the distribution of metrics for values of K in a specific range. Hence, the goal is not to find an optimal K, merely to evaluate how the algorithms perform over this range. In previous research, presented in section 3.2, this range is generally between 10 and 150, with a step size of 10. This range was also chosen here and denoted by $range(K)$.

The quantitative evaluation process is illustrated in Figure 5.3.

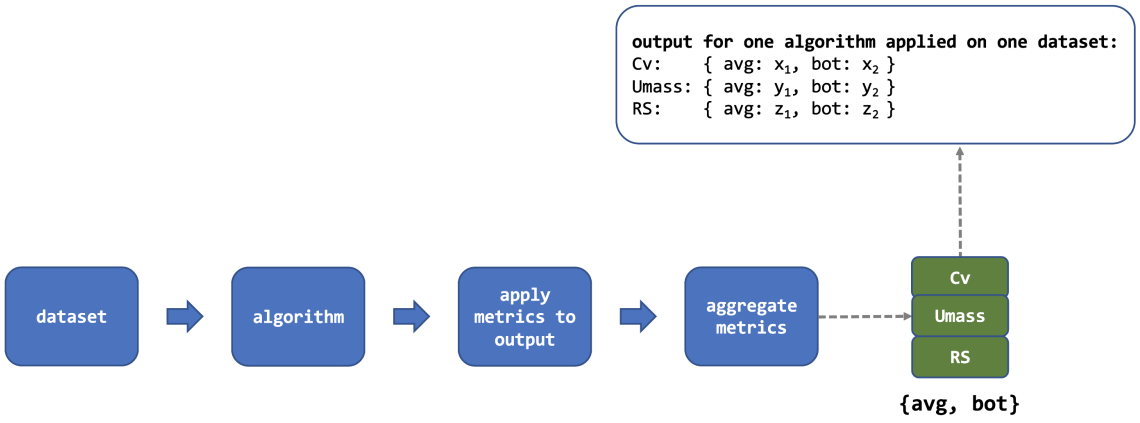

Figure 5.3: Overview of the quantitative evaluation

The quantitative evaluation process in Figure 5.3 was done two times. The first time, the goal was to evaluate and select one NMF and one LDA algorithm from the versions produced in the algorithm selection phase. The second time the goal was to evaluate these two in how well the algorithms generalized when learned with news articles from multiple news brands.

In the first step, one algorithm was evaluated on a dataset of the single-brand DN dataset type. The algorithm was rerun 15 times to each time find topics for each K in $range(K)$, where this range was 10, 20, 30, ..., 140, 150. This procedure yielded topic and topic-article assignment outputs for each of these K for one model on the single brand DN dataset. These outputs were each individually measured by the three quantitative metrics. This yielded Cv, UMass, and RS metrics for all models with K topics in $range(K)$. This procedure was then repeated for each of the algorithms from the versions produced in the algorithm selection phase.

The evaluation result of one algorithm version on the DN dataset type was thus Cv, UMass, and RS metrics, for each K topic in $range(K)$. The Cv and UMass metrics were applied to the topic output for a model learned to find K topics, and yielded a numeric vector of length K where each value of that vector gave Cv or UMass values for a particular topic. Similarly, the RS metric was applied to the topic-article assignments that a model produced for N articles and yielded a numeric vector of length N where each value of that vector gave the RS metric for a particular document that had been assigned topics by the model. The Cv, UMass, and RS metrics thus all yielded numeric vectors for one model fit to N documents to find K topics. To interpret these vectors of numbers for a model, two aggregations were applied to reduce them into the following singular value metrics.

- avg: the average of the vector
- bot: the average of the lowest 10 percent values in the vector

These aggregations were used to see the average, as well as the lower tail distribution of the metric to grasp its general stability. This bottom tail was useful to discern models with particularly poor topics in the lower end of these metrics. Cv, UMass, and RS were therefore presented by these aggregations. The practical implementation details of all three metrics are presented in sections 5.4.1 and 5.4.2.

The result of the first step was one optimal NMF and one optimal LDA algorithm, presented in more detail in section 6.3.1. These two algorithms were then evaluated in the second step, where the goal was to evaluate these two algorithms in how well they generalized when relearned with datasets of multiple news brands. In this second step, each of the two algorithms was evaluated on each of the multi-brand datasets, described in section 4.3. For each multi-brand dataset, the algorithm was rerun 15 times to each time find topics for each K in $range(K)$, where this range was 10, 20, 30, ..., 140, 150. This procedure yielded topic and topic-article assignment outputs for each K in $range(K)$ for the model on each multi-brand dataset. These outputs were each individually measured by the three quantitative metrics. This yielded Cv, UMass, and Rs metrics for all models with K topics in $range(K)$.

5.4.1 Topic Coherence

Two topic coherence metrics were used in the quantitative evaluation in this thesis, Cv and UMass. These metrics are both calculated by topics as input, where a topic is a list of terms. The Gensim library was used to measure both Cv and UMass as it provides functionality for both these metrics.

Cv was used in conjunction with the extrinsic reference corpus. This metric was thus an extrinsic topic coherence evaluation, see section 2.5.1 for detailed explanations of this metric and the importance of its extrinsic approach. The extrinsic corpus, further explained in section 4.4, was preprocessed by a slightly different procedure than the one that was applied to the corpus used to learn the model. Tokenization, lemmatization, etc. were the same, but stopwords were not removed. Also, all POS classes were kept except the class holding punctuations. The reason for these preprocessing choices was to make the corpus appropriate to use in the Boolean sliding window algorithm. The extrinsic corpus was passed to the Cv function along with the topics and produced a Cv coherence value for each topic.

UMass was used in conjunction with the intrinsic corpus, namely the corpus that was used to learn the model. This metric was thus an intrinsic topic coherence evaluation, see section 2.5.1 for detailed explanations of this metric. The intrinsic corpus was passed to the UMass function along with the topics and produced a UMass coherence value for each topic.

For a model, all its topics were passed to the Cv and UMass function respectively, which returned two lists of Cv and UMass coherence values, one for each topic. These lists were then reduced by the average and bot aggregations described earlier, giving interpretable Cv and UMass results.

5.4.2 Document-Topics Relative Sparseness

A part of the output of a topic model is a weighted numerical vector for each article, where each element corresponds to a topic and is proportional to how much of the article can be described by that topic. If a topic model is intended to partly be used to categorize articles by a single topic, it is interesting to examine how distinctly articles can be classified into one or a few of these topics. To roughly estimate this, the fraction of the largest element of the vector compared to the total sum of the vector was calculated. This yielded a score bound between zero and one, where a high score indicated that one topic described most of the article relative to all other topics, and is hence preferable if the topic model is intended to be used to classify articles. The proposed name for this metric is Relative Sparseness (RS). The RS metric produces a vector of relative sparseness values, one for each article. This vector was then reduced by the average and bot aggregations described earlier, yielding interpretable results. The grounds for using this metric was to give an approximation of how well a topic model could be used to classify articles into a single topic since it inherently

measures that property of the model as opposed to classifying articles as belonging to many topics simultaneously. However, this metric is not used in previous research on topic model evaluation and has no scientific evidence correlating with human judgment on the quality of a model. The RS metric should therefore be interpreted as a describing property of a topic model, and not as an absolute metric of the quality of the model.

5.5 Qualitative Evaluation

As stated earlier in this chapter, the qualitative judgment of topics is the most important criterion for the validity and usability of topic models. After filtering out algorithms by quantitative metrics, the remaining algorithms with corresponding models were evaluated qualitatively. The goal of the qualitative evaluation was to find one optimal topic model that yielded the most meaningful topics for Swedish news articles.

The qualitative evaluation was based on human observation and analysis. First, topics from the NMF and LDA optimal models learned from the single brand DN dataset were observed and analyzed. Second, the same procedure was done for topics from NMF and LDA optimal models learned from the multi-brand datasets. The models were assessed by highlighting important characteristics of the topics, and comparing how the topics differed semantically between the two models, as well as how they differed in how they generalized learned with news article datasets of multiple news brands. The overall quality of the respective models was evaluated, concerning differences in characteristics. This result was used as a basis for the subsequent discussion of the general viability of topic models, and for choosing the best topic model for describing Swedish news article text data, as well as the capability of topic models to describe and categorize Swedish news articles.

The qualitative evaluation was performed on six specific models, two algorithms each applied to three datasets. For all six models, the chosen number of topics K was set to 40. This naturally implicates limitations for the qualitative evaluation. However, to evaluate topics qualitatively over a $range(K)$ was seen as unreasonable within the scope of this thesis. The observations and analysis were made by the authors of this thesis, and are therefore subjective and can be questionable. The observations aimed to state the obvious characteristics of the models and leave the debatable characteristics to the subsequent discussion. This method is consistent with several previous studies. However, there are also studies conducting qualitative studies to a larger extent, having humans qualitatively evaluating topics on a large scale. In this thesis, several different topic models on different datasets could have been presented and evaluated by domain experts. This could give better grounds for comparisons of models and learning data with the aim to better understand the aspects that domain experts consider most important for using topic modeling for analysis and categorization of Swedish news articles. This was not feasible in the scope of this thesis, but the small-scale qualitative evaluation done instead was considered to give valuable results in these matters regardless.

6 Results

In this chapter, the findings from the experiments of building and evaluating meaningful topic models are presented. First, a presentation will be given of the results from the survey investigating which preprocessing choices to make, for example, part of speech classes to use. The next section presents the results from the algorithm selection phase of the NMF and LDA algorithms, suitable hyperparameters, and inference or solver methods. Then, the quantitative and qualitative evaluation results from the experiments are presented. Finally, a summary of the results is presented.

6.1 Data Preparation Results: Part of Speech Classes

The key insight from the survey was which part of speech classes that the domain experts preferred. The collected results from survey question are found in Figure 6.1 and the collected result from question two can be found in Figure 6.2. The respondents were domain experts from Bonnier News, with roles such as journalists, editors, data analysts, etc. In total, 20 answers were collected. There were three different articles for each survey answer, yielding 60 POS class evaluations in total.

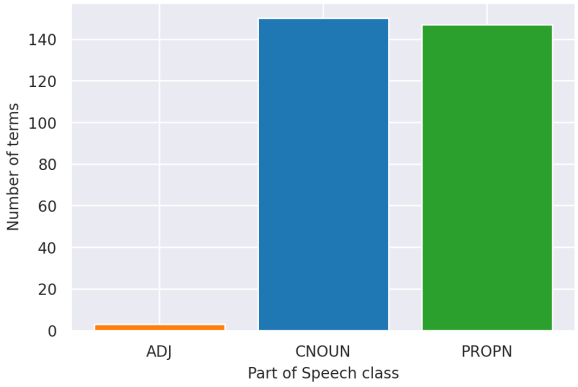

Figure 6.1: Survey result for the specified keywords per POS class

In the first question, the domain experts were asked to specify five keywords they believed described the theme of the article in the best way. This means that a total of 300 terms were collected. Figure 6.1 presents this result, where the terms are divided by their POS class. In total, the terms belonged only to three different classes, where 3 terms were adjectives (ADJ), 150 terms common nouns (CNOUN), and 147 terms proper nouns (PROPN). This result strongly indicated that the relevant POS classes in topics are common nouns and proper nouns. Between these two, there was no clear preferred option, both of them seemed to be appropriate to use in a topic according to the domain experts.

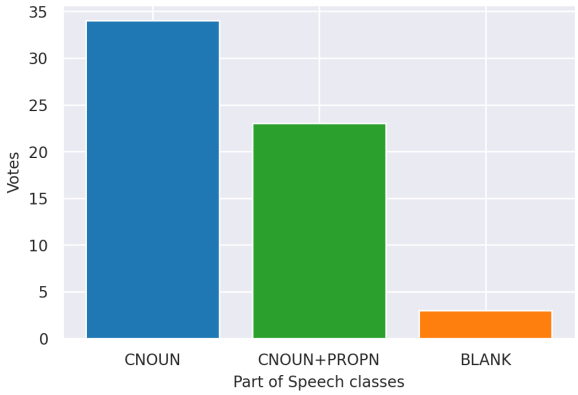

Figure 6.2: Survey result for the preferred POS class alternative

The result from the second question is presented in Figure 6.2. In this question, the domain experts were presented with two different versions of topics that the article was assigned to, one version with a topic from a model learned with a corpus that included common nouns and proper nouns (CNOUN+PROPN), and the other version a topic from a model learned with a corpus that included only common nouns (CNOUN). They were then asked to choose which topic they believed described the theme of the article in the best way. The domain experts were also given the option not to choose between the two versions, with an "I do not want to answer this question"-option (BLANK). The "common nouns only" alternative had 34 votes, the "common nouns and proper nouns" alternative had 23 votes and 3 blank answers were given. The result indicated that a topic from a model learned with a corpus only including common nouns was better to describe the articles in the survey.

The result of the survey was ambiguous. The domain experts used a lot of proper nouns when specifying keywords for articles and had a slight preference for the "common nouns only" alternative when choosing between two topics. The problems described in 5.2.1 together with the fact that the survey was answered by only 20 respondents made the result from the survey questionable. However, the result was not used standalone to make the decision which POS classes to use. The result together with theory presented in section 2.3.2, which showed that keeping the POS classes common and proper nouns were advantageous, served as a basis for the decision to filter out all POS classes except common nouns and proper nouns in the corpora used for the experiments in this thesis.

6.2 Algorithm Selection Results

The algorithm versions to evaluate were selected through the algorithm selection phase described in section 5.3. The current section will present these versions, as well as the indirect result of this phase which was an empirical understanding of how NMF and LDA hyperparameters in the scikit-learn and Gensim implementations affect model results.

An additional outcome of the algorithm selection phase was that usage of ngrams affected NMF and LDA models differently. When forming ngrams over first and last names and common phrases, the NMF algorithm tended to form topics that to a large extent contained ngrams of names. This also positively affected the quantitative scores of NMF models, most likely due to rare but statistically significant co-occurances of these names. However, topics filled with names are not deemed meaningful for describing general themes in articles. The use of ngrams, especially over named entities of multiple terms, had a negative impact on NMF in yielding meaningful topics. LDA did not favor ngrams as heavily and created meaningful topics whilst retaining the expressive power of ngrams. Ngrams was not used in the experiments in this thesis, as the Cv metric would heavily favor NMF models that produced topics heavily influenced by named entities.

6.2.1 NMF Algorithm Selection Results

The algorithm selection phase for NMF resulted in four versions, presented in Table 6.1. These NMF versions can be broken down into two parts, where the choice of the norm in the objective function is the main difference. All versions should ideally be run until convergence and the max_iterations hyperparameter should allow for this. It was set to 500 in these experiments. The initializations of the W and H matrices are set by the init hyperparameter. The optimal settings were found to be nonnegative double singular value decomposition (NNDSVD) for coordinate descent (CD) solver, and nonnegative double singular value decomposition with zeros filled with the average of X (NNDSVDa) for the multiplicative update (MU) solver, for both methods to initially update optimally.

NMF Optimized for the Frobenius Norm

These NMF versions minimize the Frobenius norm in the objective function and both the CD and the MU solver can be used for this norm. CD achieved similar or marginally better quantitative results compared to MU, but the MU solver was significantly faster. The Frobenius norm models produced topics of relatively high coherence scores, but with relatively poor results in assigning specific topics to articles.

Regularization can be applied with the Frobenius norm. Only the L2 norm gave meaningful results, as an influence by the L1 norm caused broad and bland topics. The magnitude of the regularization effect is tuned by the hyperparameter alpha. High alpha led to an increase in coherence scores and stability over many topics, but significantly reduced the potential of assigning specific topics to articles. Inappropriately large alpha values had the effect that some topics merged and became almost identical. An alpha value of 1.5 was found as a suitable maximum in the algorithm selection phase. It should be noted that the effect alpha has on the algorithm appears to be correlated with the dataset size, which in this case was 10 000 articles. Both the MU and CD solver can be used with regularization.

NMF Optimized for the Kullback-Leibler Divergence

These NMF versions minimize the Kullback-Leibler divergence in the NMF objective function. Only the MU solver can be used for this objective function. The models that were optimized for this norm were characterized by poorer coherence scores compared to the Frobenius norm but achieved significantly better scores in assigning specific topics to articles. Regularization was found to have minimal effect in this objective function.

6.2.2 LDA Algorithm Selection Results

The algorithm selection phase for LDA resulted in two versions, presented in Table 6.2. These two versions are implemented differently, one with the Gensim built-in algorithm, and the other with the Java-based package MALLET. Both versions were optimized by its hyperparameters. The main difference is how the two versions perform inference. Gensim version utilizes the inference technique online variational Bayes, and MALLET version utilizes the inference technique Gibbs sampling. These are explained further in section 2.4.3, but notable here is that online variational Bayes turned out to be significantly faster compared to Gibbs sampling in the experiments conducted in this thesis.

LDA Gensim

Two important hyperparameters for the Gensim version are alpha, the Dirichlet hyperparameter for the document-topic density, and eta, the Dirichlet hyperparameter for the term-topic density. These two hyperparameters were found to produce the models with the most meaningful topics when they were set to auto, meaning the algorithm learns asymmetric priors from the corpus. The algorithm should ideally be run until convergence, in the experiments this means that the optimal hyperparameter passes were shown to be 3 and the hyperparameter iterations were shown to be 100.

LDA MALLET

The hyperparameters for the MALLET version were via experiments shown to be optimal with the default values. The hyperparameters in focus in the experiments were the number of iterations and alpha. Increasing the number of iterations did not result in better coherence or more meaningful topics when manually inspecting them, neither did decreasing this number. Experiments tuning the alpha hyperparameter resulted in that the default value of 50 was a good choice.

6.2.3 Topic Model Algorithm Versions

The algorithm versions that the algorithm selection phase resulted in is presented in Table 6.1 and 6.2.

Table 6.1: The four NMF algorithm versions selected in the algorithm selection phase

Hyperparameter	**Frob MU non-R**	**Frob CD non-R**	**Frob MU R**	**KL**
n_components (K)	range(K)	range(K)	range(K)	range(K)
beta_loss	Frobenius	Frobenius	Frobenius	Kullback-Leibler
solver	MU	CD	MU	MU
max_iter	500	500	500	500
init	NNDSVDa	NNDSVD	NNDSVDa	NNDSVDa
alpha	0	0	1.5	0
l1_ratio	0	0	0	0

Table 6.2: The two LDA algorithm versions selected in the algorithm selection phase

Hyperparameter	**Gensim**	**MALLET**
num_topics (K)	range(K)	range(K)
passes	3	-
iterations	100	1000
alpha	auto	50
eta	auto	-

6.3 Quantitative Evaluation Results

This section presents the results from the quantitative evaluation of the NMF and LDA algorithms by the Cv, UMass, and RS metrics. First, the results from the evaluation of NMF and LDA models learned by the single-brand DN dataset are presented. The two optimal model's corresponding algorithms from this evaluation are then applied to multi-brand datasets, to evaluate how these algorithms generalize when learned with more heterogeneous data.

6.3.1 NMF and LDA Algorithm Comparisons

The algorithm versions listed in Table 6.1 and 6.2 were applied on the DN dataset and evaluated by the quantitative metrics. The purpose of this procedure is to select one optimal NMF algorithm and one optimal LDA algorithm.

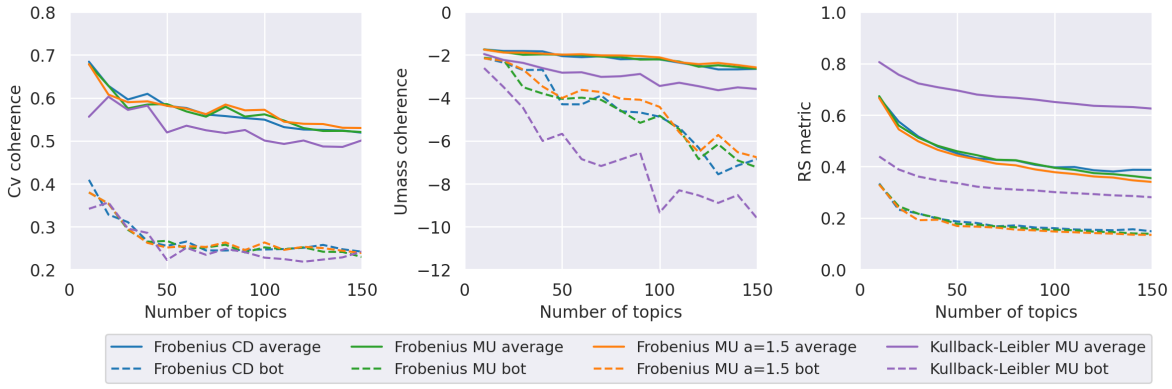

Figure 6.3: Quantitative results for NMF algorithms applied on the DN dataset

The NMF results are presented in Figure 6.3. A clear declining trend was observed for all four learned NMF models as the number of topics K increased. The Cv trend was not monotonic in its decrease however and had local variational spikes for small ranges of K, indicating that there can be a local optimum of a number of topics to find in small local spans of K. The model optimized for the Kullback-Leibler norm performed notably worse in coherence metrics compared to its peers. It markedly outperformed the others in distinctly assigning a single topic to an article on average as shown by the RS metric. Among the models optimized with the Frobenius norm, the MU and CD solvers appeared to give similar results on all three metrics. The Frobenius regularized model achieved marginally higher coherence scores but performed marginally worse in the RS metrics. The models behave similarly for the bottom 10 percent aggregation (dashed line) and these bot metrics are notably worse in general compared to the average, for both Cv, UMass, and RS. Although the regularized Frobenius model performed slightly better in coherence scores, its tendency to create blander topics, as discovered in the algorithm selection phase, led to uncertainty for using this model. By these criteria, the NMF algorithm that optimizes the Frobenius norm with the MU solver and without regularization was chosen as the optimal NMF algorithm version.

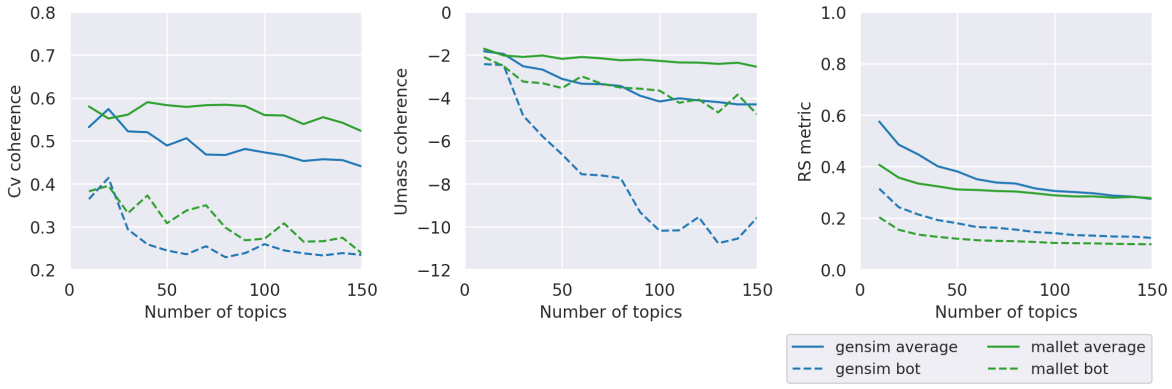

Figure 6.4: Quantitative results for LDA algorithms applied on the DN dataset

The LDA results are presented in Figure 6.4. The increase in the number of topics led to a similar decreasing coherence trend for the Gensim model. On the contrary, MALLET peaks in average coherence between 40 and 80 topics. The coherence scores for the model learned by the Gensim version declined with increasing topics, especially so for UMass. The MALLET model showed stability in both these metrics as the number of topics increased. The RS metric decreases for both with increasing topics. There was a significant gap between the average score and the bottom 10 percent score for both LDA models, but this was not as pronounced as for NMF. The results made it clear that the MALLET model was superior in producing coherent topics in comparison to the Gensim model. The MALLET algorithm was chosen as the optimal LDA algorithm version.

In general, the NMF algorithms performed better or at least as good as the LDA algorithms on average, especially on fewer topics. The NMF algorithms all had clear declining trends in coherence score as the number of topics K increased. This declining trend was also observed for LDA, especially in UMass and RS, but it was not as pronounced as for NMF.

When comparing the optimal NMF and optimal LDA model it was even more clear, compared to the general case, that the optimal LDA did not have this notable declining trend. The optimal LDA achieved stable coherence scores for any number of topics up to 100 but had slightly decreasing coherence scores from 100 to 150 topics. Although the optimal NMF model was superior in the Cv coherence score to the optimal LDA model for the lower end of the topic spectrum, they both had approximately the same coherence score for the higher end of the spectrum. For UMass these two versions were alike. The optimal LDA did perform better in the bottom 10 percent aggregation for both coherence scores, however, and in particular for UMass compared to the optimal NMF model. For the RS score, the optimal NMF exceeded the optimal LDA on average, while the bottom 10 percentile was alike.

This result made it clear that the optimal NMF model generally performed better or as good as the optimal LDA model on average for all scores, but performed on par or worse in

coherence for its most incoherent topics. NMF was judged to be better but less stable than LDA in these metrics. The RS metric also showed a slight advantage for NMF for assigning articles to a single topic.

6.3.2 NMF and LDA Generalizability

The optimal NMF and LDA algorithm versions from the previous section were applied on the DN/Di/HD and the BN datasets, presented in section 4.3, to respectively learn new models over multiple brands. These NMF and LDA models were then evaluated by the quantitative metrics as described in section 5.4. The purpose of this procedure was to evaluate how these algorithms generalized when learned with news articles datasets from multiple news brands and how the quantitative metrics differed between these datasets.

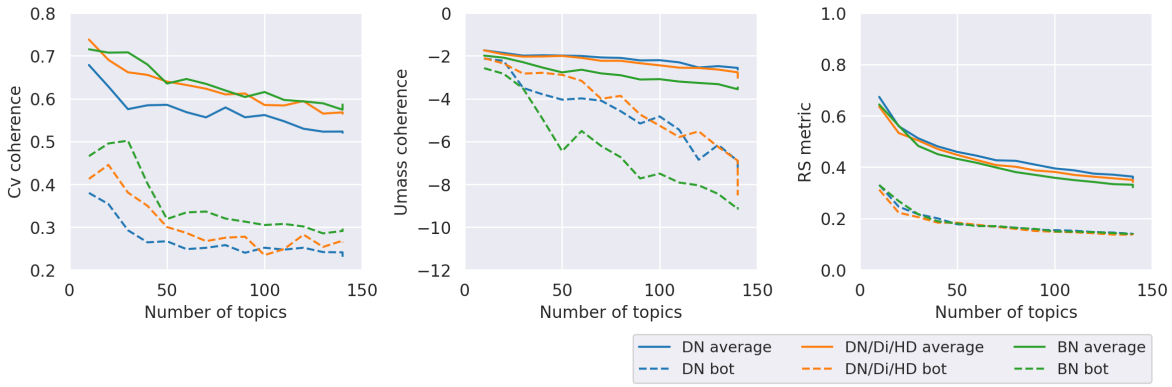

Figure 6.5: Quantitative results for the NMF algorithm applied on the three datasets: DN, DN/Di/HD and BN

The NMF results for generalizability over several news brands are presented in Figure 6.5. The NMF models learned with the DN/Di/HD and BN datasets showed similar trends in all scores compared to the optimal NMF model learned with the DN dataset. The models learned by these multi-brand datasets performed significantly better in the Cv coherence metric compared to the single-brand DN dataset, which was somewhat unexpected. For the UMass coherence metric, the NMF models learned with the DN dataset performed similarly to the model learned with the multi-brand DN/Di/HD dataset. However, the model learned with the multi-brand BN dataset performed marginally worse. All three models learned by each of the three datasets performed similarly in the RS metric. Despite the minor differences in coherence metrics, by observing the trends in the results it was concluded that the NMF algorithm, in a quantitative sense, generalized well when learned with news articles from multiple news brands.

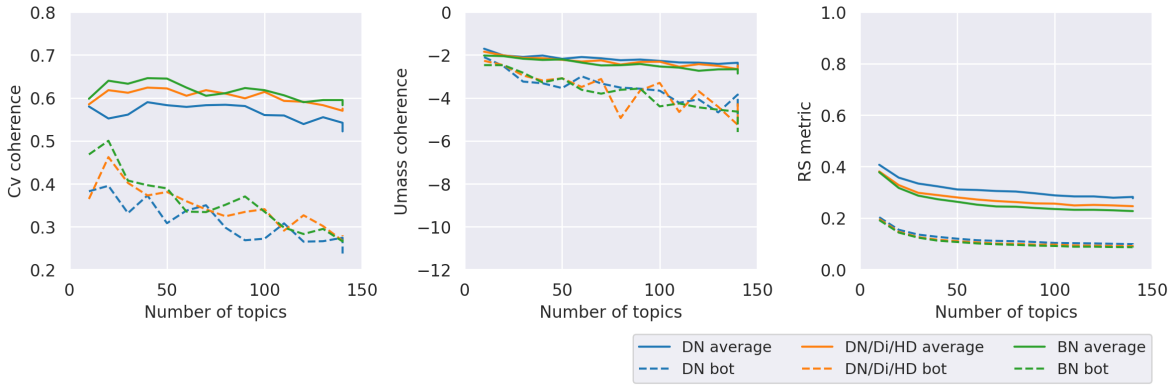

Figure 6.6: Quantitative results for the LDA algorithm applied on the three datasets: DN, DN/Di/HD and BN

The LDA results for algorithm generalizability over several news brands are presented in Figure 6.6. Likewise to NMF, the LDA models learned with the multi-brand DN/Di/HD and BN datasets showed similar trends in all scores compared to the LDA model learned with the single-brand DN dataset. Similarly to the NMF results, the LDA models learned with the multi-brand datasets generally performed better than the LDA model learned by the single-brand DN dataset for Cv coherence. The LDA models learned by each of the three datasets performed similarly in UMass coherence metrics as well as in the RS metric. By these trends, it was concluded that the LDA algorithm also, in a quantitative sense, generalized well when learned with news articles from multiple news brands.

Both the NMF and the LDA algorithms generalized well to datasets of multiple news brands in a quantitative sense. When comparing the NMF and LDA algorithms in their ability to generalize over news articles spanning different news brands, there were no major differences. One thing worth mentioning is that, for the extrinsic Cv coherence metric, the models learned with the multi-brand datasets performed better compared to the single-brand DN dataset. A reason for this could be that the extrinsic dataset, used as a reference corpus in the calculations of this metric, consisted of articles from multiple brands, and thus better validated the topics in the multi-brand NMF and LDA models. The reason could potentially be that the topics from these models are better mirrored in the extrinsic data. Still, there were no overlapping articles between the datasets and the extrinsic data, and the composition of the extrinsic data was motivated by its ability to capture the Swedish language within news media in a general sense. In section 4.4 a detailed description of the extrinsic data is given.

6.4 Qualitative Evaluation Results

This section presents the results of the qualitative evaluation of the NMF and LDA algorithms. First, a selection of topics from the NMF and LDA optimal models learned with the single-brand DN dataset are showcased. Next, a new selection of topics is presented from NMF

and LDA models learned by the optimal algorithms but applied on the different multi-brand datasets. This new selection of topics is showcased to compare NMF and LDA to each other, but also to show how the topics from the respective algorithms change when the algorithms are applied to different datasets. This gives a qualitative perspective on how the topics differ semantically as the algorithms are applied on multiple news brands. These topics are also chosen as illustrative examples. A topic is presented as the five most heavily weighted terms in the topic, as described in section 2.4. A full list of all topics for all qualitatively evaluated models can be found in Appendix C, with English translations.

6.4.1 Topics from DN Data

Both the NMF and LDA algorithms were each respectively applied on the DN dataset to find 40 topics. The number 40 was chosen since the LDA and NMF algorithms both produced relatively high quantitative metrics for this number. For presentability reasons, a selection of the topics is presented since they highlight characteristics for the two models that were found important. NMF topics are presented in Table 6.3. All 40 topics can be found in Appendix C. The selected LDA topics are presented in Table 6.4, with all 40 topics in Appendix C.

Table 6.3: Selected NMF Topics from DN Data

A1. barn förälder familj mamma sverige
A3. trump president donald usa demokrat
A4. may brexit storbritannien eu theresa
A5. match mål lag spelare seger
A6. kina coronavirus virus usa wuhan
A9. iran usa irak soleimani attack
A16. thunberg greta klimataktivist klimat värld
A20. patient vård läkare region sjukvård
A24. öberg vm karlsson frida johaug
A38. sverige asap rocky norge artist
...

A clear observation for the NMF model was that proper nouns (i.e. named entities) such as persons or countries can get intertwined with concepts if a person or country is very common in the context of a certain concept. For example, topic A16: *thunberg, greta, klimataktivist, klimat, värld (thunberg, greta, climate activist, climate, world)* merges the concept of climate

activism with the person Greta Thunberg. This is reasonable because Greta Thunberg is often mentioned in articles on climate activism, or vice versa, and the model successfully recognizes this pattern. This also means that articles that to some extent are about climate activism but not about Greta Thunberg, in effect have a high chance of being classified as topic A16 anyway, regardless if it contains her name or not. This proper noun connection was not the case for all topics though, as some topics were created only by common nouns and reflected concepts rather than persons. For example, topic A20: *patient, vård, läkare, region, sjukvård (patient, care, doctor, region, healthcare)* is clearly about healthcare, without proper nouns in it.

Another recurring phenomenon related to proper nouns was that the NMF model tended to form topics mostly out of people's names. Topic A24: *öberg, vm, karlsson, frida, johaug (öberg, world championship, karlsson, frida, johaug)* is made up of four names that are related since they are often mentioned in similar articles. Topic models create topics of terms that co-occur frequently which makes this topic reasonable, and it can still achieve high coherence scores as these metrics are also based on term co-occurrences. However, a topic of only names can be argued not to be meaningful, a direct effect of using proper nouns.

Most topics were considered to be relatively timeless in the sense that the concepts they represent can be thought to persist through time, such as topic A20 on healthcare. However, certain news events could themselves create topics if the event was covered to enough extent in the data. Topic A9: *iran, usa, irak, soleimani, attack (iran, usa, iraq, soleimani, attack)* most likely relates to the US military attack on Iranian general Qasem Soleimani in 2020. This event is particularly time-specific but received enough coverage in the data, which is the only criteria for the topic model. Another example is topic A4: *may, brexit, storbritannien, eu, therese (may, brexit, britain, eu, theresa)* that relates to British politics and the Brexit event.

Table 6.4: Selected LDA Topics from DN Data

D4. svar fråga samhälle insändare sverige
D6. kina land hongkong virus coronavirus
D7. match mål lag period säsong
D12. stockholm stad göteborg kommun plats
D16. kyrka häst kläder namn väg
D17. värld tal greta historia thunberg
D20. sverige land utsläpp klimat projekt
D30. iran usa ryssland turkiet israel
D36. sjukhus vård region patient läkare
D39. bok författare roman liv berättelse
...

The phenomenon that people or places get tied together with concepts was also prominent in topics created by the LDA model. Topic D17: *värld, tal, greta, historia, thunberg (world, speech, greta, history, thunberg)* ties together Greta Thunberg with speech and history, which can be interpreted as a concept different from climate activism. LDA topics also had a similar tendency to form topics of only proper nouns that co-occur frequently, as in topic D30: *iran, usa, ryssland, turkiet, israel (iran, usa, russia, turkey, israel)*.

An interesting observation in the LDA topics, and for topic models in general, was that topics can allude to different types of concepts. For example, topic D12: *stockholm, stad, göteborg, kommun, plats (stockholm, city, gothenburg, municipality, location)* refer to geographic locations in Sweden and places in general, while topic D4: *svar, fråga, samhälle, insändare, sverige, (answer, question, society, letter to the editor, sweden)* refers to a type of articles and general societal questions. Topic D12 refers to geographic locations and topic D4 to a type of article. These are under the same taxonomic system but highly differ in the type of concept they refer to.

LDA generally produced semantically coherent topics, but with some exceptions. Topic D16: *kyrka, häst, kläder, namn, väg (church, horse, clothes, names, road)* does not appear to have any clear, underlying semantically interpretable concept. In contrast, other topics offer clear semantic interpretation like topic D39: *bok, författare, roman, liv, berättelse (book, author, novel, life, story)* which represent the concept of books and stories.

In contrast to one another, NMF and LDA produced similar types of topics. With the DN dataset, the two algorithms created near-identical topics (i.e. topic A5 for NMF and topic D7 for LDA). Through the experiments, there was some indication that NMF tended to produce some topics more related to specific events than LDA. Two examples of such specific events can be seen in the NMF topics, where topic A9 refers to the US attack on Iranian Soleimani, and topic A38 appears to refer to artist Asap Rocky and his visit to the Nordics (which received a lot of media attention). These events are not present in the LDA topics, and no clear "time-specific events" can be seen in LDA topics in general. Another contrasting factor is the number of proper nouns in NMF topics compared to LDA. When counting proper nouns of the top five terms in each topic for the 40 topics produced on the DN dataset by each model, NMF produced 83 proper nouns and LDA 44 proper nouns out of 200 total terms for each model.

To conclude, LDA tended to be more stable and produced broader, more general topics than NMF, with the drawback that these topics might be too unspecific for representing meaningful themes in articles. NMF on the other hand tended to produce more specific topics, related to specific events, names, concepts, etc.

6.4.2 Topics from Multi-Brand Data

The optimal NMF and LDA algorithms were then used to learn models with the DN/Di/HD and the BN datasets, and the resulting topics were evaluated to qualitatively assess multi-brand generalizability. Both algorithms were applied on these datasets respectively, to find 40 topics in each case. The same procedure was carried out for the LDA algorithm.

A selection of the NMF topics from the DN/Di/HD dataset is presented in Table 6.5, with all topics in Appendix C. They are chosen due to the reason that they highlight the most important characteristics of the model.

Table 6.5: Selected NMF Topics from DN/Di/HD Data

B1. aktie bolag krona miljon stockholmsbörs
B2. trump president donald usa demokrat
B6. match mål lag säsong poäng
B13. kina coronavirus virus tull handelskrig
B14. kvinna våldtäkt våld brott misshandel
B16. bil elbil förare väg tesla
B28. olycka sjukhus ambulans väg lastbil
B34. börs index wall street dow
B36. fat oljepris lager vecka olja
B37. may theresa parlament brexit premiärminister
...

A selection of the NMF topics from the BN dataset is presented in Table 6.6, with all topics, and in Appendix C.

Table 6.6: Selected NMF Topics from BN Data

C3. match mål lag poäng seger
C6. bil olycka räddningstjänst väg ambulans
C9. kina hongkong usa land coronavirus
C10. barn förälder förskola familj mamma
C11. trump president usa donald demokrat
C26. byrå kund kampanj varumärke resumé
C27. skövde skara ifk skaraborg falköping
C28. bok författare roman liv berättelse
C39. greta thunberg klimat värld klimataktivist
C40. smhi län jönköping temperatur varning
...

There were a notable number of similar topics from the NMF models learned by the multi-brand datasets in comparison to the NMF model learned by the single-brand DN dataset. There is a topic about US politics and Donald Trump in all three NMF topics-sets, corresponding to topic A3 in the DN dataset, topic B2 in the DN/Di/HD dataset, and topic C11 in the BN dataset. Similarly, a topic about sports was present in topic A5 in the DN dataset, topic B6 in the DN/Di/HD dataset, and topic C3 in the BN dataset. These were only a few examples among many and showcased notable similarity in taxonomy between the NMF models learned with the different datasets. There were also clear differences between these NMF models.

First, the span of concepts which the topics covered differed from between the models. This is best illustrated by topics related to economics, finance, or money. The NMF models learned by the DN dataset has two topics related to these concepts (topics A14 and A25), the DN/Di/BN dataset model has fourteen (topics B1, B7, B9, B10, B11, B15, B18, B22, B24, B26, B27, B33, B34, and B36) and the BN dataset model has four such topics (topics C4, C16, C24, and C29). Dagens Industri (Di) is an industry magazine on business and makes up 33% of the articles in the DN/Di/HD dataset, a small fraction of the BN dataset, and is not present in the DN dataset. The presence of that brand in the datasets determines how many topics are formed and somewhat related to economy, finance, and money. The model learned with the DN/Di/HD dataset thus yielded unique, more granular topics on economics and finance, such as topic B34: *börs, index, wall, street, dow (stock exchange, index, wall, street, dow)* which is clearly about stocks, and topic B36: *fat, oljepris, lager, vecka, olja (barrel, oil price, storage, week, oil)* which is clearly about oil prices. Similarly, the NMF models learned with the DN and BN datasets both yielded several topics each on healthcare and accidents, whilst there were fewer such topics from the model learned with the DN/Di/BN dataset. The NMF model learned with the BN dataset showed more topics related to local matters such as weather (topic C40) or local places (topic C27).

Second, the dataset content determines the intra-topic semantics. For example, topic B13 in the DN/Di/HD dataset: *kina, coronavirus, virus, tull, handelskrig (china, coronavirus, virus, customs, trade war)* relates China with the covid19 pandemic virus, but also with international trade. There is also a topic about China from the NMF model learned with the DN dataset (topic A6), but this topic has no clear connection to the concept of international trade. The China topic (topic C9) from the NMF model learned with the BN dataset also shows no connection to international trade.

To summarize, the NMF algorithm was judged to generalize well in a qualitative sense when applied on heterogeneous datasets of articles of multiple brands, but that the dataset contents heavily determine the topic output. Next, a selection of the LDA topics from the DN/Di/HD dataset is presented in Table 6.7, with all topics in Appendix C.

Table 6.7: Selected LDA Topics from DN/Di/HD Data

E1. musik låt artist band scen
E2. storbritannien avtal eu brexit johnson
E3. mat restaurang djur häst vin
E5. vård patient studie sjukhus läkare
E6. månad vecka januari antal februari
E7. match mål lag säsong period
E21. sverige fråga samhälle svar problem
E24. kina usa dollar miljard tull
E35. bil volvo krona fordon väg
E36. vatten tåg väg skog område
...

A selection of the LDA topics from the BN dataset is presented in Table 6.8, with all topics in Appendix C.

Table 6.8: Selected LDA Topics from BN Data

F4. vatten skog plan område grad
F5. match lag säsong mål spelare
F13. vecka månad sommar januari slut
F14. skövde skara förening falköping skaraborg
F19. usa trump kina president land
F23. jönköping län eksjö nässjö värnamo
F26. stad hus område plats lokal
F31. mat restaurang jul kött vin
F36. fråga problem samhälle exempel svar
F38. sverige land antal värld svensk
...

The LDA models also showed trends of generalizability across multiple datasets, as a number of topics became nearly identical for the learned models over the three different datasets. When observing all 40 topics from the three LDA models respectively there were topics between the datasets that related to similar concepts. For example a topic about sports, corresponding to topic D7 in the DN dataset, topic E7 in the DN/Di/HD dataset and topic F5 in the BN dataset. Similarly, a topic about a type of articles and general societal questions was present in topic D4 in the DN dataset, topic E21 in the DN/Di/HD dataset, and topic F36 in the BN dataset.

The LDA model that was learned from the DN/Di/HD dataset also produced more topics related to economic concepts compared to the LDA models from the DN and BN datasets. The DN dataset model yielded two such topics (topic D18 and D23), the DN/Di/HD dataset model yielded six (topic E10, E18, E20, E22, E29, and E33) and the BN dataset model yielded two (topic F1 and F32). This is notably less economic topics from the DN/Di/HD dataset compared to the NMF model. The LDA economic topics were also less granular and had no specific topics relating to the stock exchange, or oil prices for example.

LDA also showed tendencies to modify intra-topic semantic meaning depending on the data it was learned with. For the DN/Di/HD dataset (with a large part economic articles), topic E24: *kina, usa, dollar, miljard, tull (china, usa, dollar, billion, customs)* relates China with terms related to money and trade (although different terms than for the NMF topic from the same dataset). The LDA model learned with the DN dataset creates the China-related topic D6: *kina, land, hongkong, virus, coronavirus (china, country, hongkong, virus, coronavirus)*, and the LDA model learned with the BN dataset creates no direct China topic.

In general, LDA models showed tendencies to create less specific topics with respect to the data it was learned with, compared to NMF which tended to create topics out of specific events or concepts related to specific segments in the data. LDA tended to create topics that broadly fit the whole dataset, with less regard to specific patterns in smaller segments in the data. Examples of such more general topics are topic E6: *månad, vecka, januari, antal, februari (month, week, january, number, february)* and topic E36: *vatten, tåg, väg, skog, område (water, train, road, forest, area)* in the DN/Di/HD dataset model, or topic F26: *stad, hus, område, plats, lokal (city, house, area, place, local)* and topic F38: *sverige, land, antal, värld, svensk (sweden, country, number, world, swedish)* in the BN dataset model. No similar topics with broad semantic interpretation could be found in the NMF models.

The main findings from the qualitative inspection of the NMF and LDA topics in this section were that both algorithms could create topic models with meaningful topics when applied to datasets of news articles from multiple brands.

6.5 Summary of Results

To illustrate the main findings in this chapter, the key takeaways are summarized below.

Data preprocessing was instrumental to the results of the topic model
Proper data cleaning and feature selection had a particularly high impact on the quality of the topic models in the experiments. Regarding the part of speech classes, proper nouns and common nouns were found to be most descriptive of topics, as specified by news media domain experts. When given the choice between common nouns only or both, the experts showed an inclination towards topics with only common nouns. Both types of nouns were used in the experiments in this thesis. This led to some meaningless topics that consisted of only proper nouns. All evaluated models created such topics to some degree. Proper nouns such as names of people or places were often found to be tied to concepts in topics, and concepts became tied to names, like topics on climate activism and Greta Thunberg.

The inter-topic distribution and intra-topic semantics was determined by the dataset
The distribution of content in articles in the dataset determined the span of concepts that the topics covered, how many topics were created on related concepts and how granular these topics became, as well as what concepts were related inside topics. The topic models created by a dataset with niched articles produced many granular topics in that niche, while models learned with broader datasets spanned a substantially wider range of different concepts with more general topics. The contents of the datasets created connections between related concepts, such as between China and international trade, or China and the covid19 pandemic.

Topics alluded to different types of concepts
Since topics are created only from statistical term co-occurrences, they can have different meanings. In the results, topics could allude to general concepts such as healthcare, whilst others described a certain event or to the concept of a geographic location. There was little semantic conformity in the taxonomic system created by the topic models, combined with a general high variability of semantic interpretability of all topics from a single model.

NMF and LDA showed different strengths and weaknesses
The results showed that NMF was generally better in quantitative scores on average but showed a decreasing trend as the number of topics increased. LDA performed markedly worse for fewer topics but held similar metric scores as the number of topics increased. The lower 10 percentile in the quantitative metrics were significantly worse than the average for both methods, but LDA showed less variance compared to NMF. The optimal algorithm version was selected as the non-regularized, Frobenius optimized NMF algorithm with the

MU solver and the Gibbs sampling MALLET implementation of LDA with auto-optimized hyperparameters. In a qualitative sense, NMF tended to form more granular, specific topics related to specific patterns in the dataset, whilst LDA tended to create broader, sometimes semantically incoherent, topics. NMF topics also favored proper nouns more heavily in topics compared to LDA.

NMF and LDA differed in how they generalized to several news brands

Both NMF and LDA generalized well when learned with news articles from multiple news brands in a quantitative sense. By qualitative observations, NMF indicated a higher sensitivity to distinct patterns in the dataset and formed more granular topics related to similar concepts to match those patterns, but completely unrelated to other parts of the dataset. On a dataset of 33% economic, business, and financial-related articles, NMF created many granular topics related to those concepts, but these were mostly unrelated to the remaining 66% of the dataset. LDA created fewer such topics and more general topics related to the holistic content of the dataset.

It was difficult to find an objectively optimal topic model

Both NMF and LDA showed different strengths and weaknesses, advantages, and limitations. Determining one best model in general, comprising both quantitative and qualitative aspects, was not feasible by the experiments in this thesis. Both algorithms have their respective use cases depending on the data and questions one attempts to answer by the topic modeling approach.

7 Discussion

In this chapter, the results from the experiments in this thesis are discussed. The thesis aimed to explore the application of automated topic generation by the unsupervised machine learning method topic modeling of Swedish news article text data. Different such models were applied to news article datasets of different news brands and evaluated by quantitative metrics and qualitative human judgment, to investigate the potential of topic modeling of Swedish news articles. Further, the aim was to evaluate if topic modeling is a valid method to use as a uniform categorization framework for Swedish news articles. Supported by this context, the purpose of this thesis was to provide an understanding of the validity, the viability of use, and limitations for such a topic model categorization framework for news articles from one brand or spanning multiple brands. The posed research questions in section 1.1 will be answered by discussing the results of the experiments in this thesis in relation to previous research that was presented mainly in chapter 3 and to some extent in chapter 2.

Which topic model algorithm and data preparation yields the most meaningful topics of Swedish news articles within one news brand?
A major realization from the experiments is that, given a dataset, data preparation is paramount to the success of topic modeling in finding coherent topics and for accurately assigning articles to them. The aspects of data preparation can be broken down into three parts: dimensionality reduction, feature selection, and data representation.

Prior dimensionality reduction techniques on the input corpus were found to greatly affect the topic model quality and processing speed. All of these findings are not explicitly stated in the results since many were found through the exploratory work of selecting the NMF and LDA algorithm versions. Stopword removal and to filter out infrequent terms led to significantly increased performance in computations and less noisy topic outputs, as suggested by previous research (Sarkar 2019). Lemmatization was found to yield better results compared to stemming, with more interpretable terms and a better ability to reduce dimensionality. This is reasonable due to the complex nature of the Swedish language as explained in section 4.2 and confirms that Germanic languages benefit from lemmatization compared to stemming as suggested by Haselmayer and Jenny (2014). Lemmatization does lead to less human-readable terms in topics since it is the lemma that is presented, as opposed to the original term, but its benefits arguably outweigh this downside.

Feature selection by selecting part of speech (POS) classes to keep in the datasets played a key role in the experiments. The survey clearly indicated a domain expert preference to use nouns (common and proper) as keywords for describing articles. Nouns were also the most common class in topics from models that were learned with corpora containing all POS classes. This confirms previous research (Martin and Johnson 2015; Jacobi, Van Atteveldt, and Welbers 2016) in that nouns are the most relevant POS class in topics, both from the perspective

of human domain experts and by topic modeling algorithms in themselves. However, the results of this thesis indicate some complications with using proper nouns in topic modeling applications.

Firstly, proper nouns represent names and these get tied together with concepts in topics, like the name Greta Thunberg and the concept of climate activism. This clearly illustrates a statistical significance between the name and the concept but does imply some complications for creating a generalized framework for categorizing articles. Given a topic with a certain concept and a certain name in it generated from a topic model learned on a sufficiently large corpus, then there will undoubtedly be articles in the corpus that are about the concept but unrelated to the name, and vice versa. Classifying articles correctly to the concept in the topic but incorrectly to the name in the topic can be highly misleading and detrimental to a categorization framework. Secondly, the experiments indicated that some topics tended to be formed by proper nouns only. This is also problematic as a topic that only consists of names are not considered to be meaningful in the sense that the topic does not directly represent a concept. These are the two main issues with using proper nouns to represent meaningful topics for categorizing articles. There were topics in the results that contained no proper nouns, such as topics that were clearly about healthcare. This implies that it can be advantageous to use only common nouns for topic models, and is also strengthened by the slight inclination for common nouns only in the survey results.

Another powerful feature selection method that warrants consideration is ngrams, as advocated by Mikolov et al. (2013). This method was not used in the experiments. It was discovered that when applying the ngram options directly on the input corpus, unwanted ngrams tended to be formed, mainly that connected titles and names such as *president obama* with little meaningful value, while missing other valuable ngrams such as *black lives matters*. To account for this, a controlled ngram option was proposed which formed predefined ngrams (such as *black lives matters*) and ngrams over names (such as *barack obama*) by utilizing named entity recognition (NER) tags extracted from the corpus. However, this option was not used either since the NMF algorithm tended to create topics that mostly consisted of ngram names which is not considered as a meaningful topic while gaining an increase in coherence scores. The predefined ngram method did not affect LDA in the same negative manner and instead enhanced the expressiveness of LDA models. The reason for this could be because of the tf-idf representation that is used by NMF and not by LDA. Ngrams can thus be a valuable data preparation method, but it can lead to unwanted effects depending on the method used and the desired results. Due to variability in how ngrams affected NMF and LDA, they were not part of the evaluation of the algorithms. It should be noted that most of the ngram related issues in the experiments arose on phrases with proper nouns, indicating that a common nouns only approach might benefit more from this feature selection method.

The data representation for both NMF and LDA is in the form of bag-of-words but differs in the sense that NMF uses a tf-idf weighting in the document-term matrix, while LDA uses an unnormalized term frequency weighting in this matrix. The tf-idf representation weighs unusual terms higher which adds more prior information compared to the simpler term frequency counts for LDA, as noted by Chen et al. (2019). News articles contain a high amount of diverse names, especially so in corpora of news articles of multiple brands of diverse areas. Names, and especially ngram-ed names formed over first and last names, will be rare in the corpus and thus gain high tf-idf values. This is a likely reason why NMF produced almost twice as many proper nouns in its top-five term topics in the qualitative evaluation results, and why ngrams formed over first and last names led to unwanted results and had a negative impact for NMF.

The previous research presented in chapter 3 indicated that LDA generally yields more coherent and stable topics compared to NMF (Stevens et al. 2012; M'sik and Casablanca 2020; Suri and Roy 2017). The quantitative results in this thesis confirm that LDA models learned by Gibbs sampling are more stable in topic coherence compared to NMF. On average, however, NMF outperforms LDA in a quantitative sense for coherence, especially so for models of few topics. This contrasts with previous research, but it should be noted that those studies used English articles. Whether the findings in this thesis are related to the more complex nature of the Swedish language (explained in section 4.2) as opposed to English is out of the scope of the thesis, but this is a possibility. Furthermore, previous research suggests that NMF has a better ability to be used to classify articles into distinct topics, which is also suggested by the findings in this thesis through the RS metric. This metric has no grounds in previous research, but it is found to describe an important characteristic of topic models when it is desirable to describe articles by a single topic.

Previous research also favors LDA models when topics are intended to be presented to humans as they are considered to be more meaningful and interpretable (Stevens et al. 2012; M'sik and Casablanca 2020; Suri and Roy 2017). This is a simplified and somewhat shallow conclusion. The qualitative results in this thesis offer a deeper analysis of the differences between NMF and LDA topic models. A major finding is that NMF tends to find topics that represent specific, distinct patterns in segments of the data. LDA has a slightly different tendency to generate holistic topics that can better be applied in the whole dataset. LDA is thus less likely to generate topics that are only applicable to patterns that only exist in specific segments in the data. In news data, segments can be specific events or granular concepts in a subset of articles in the corpus, contrasted to holistic concepts that are more likely to be present in all articles to some extent. On the dataset with a high fraction of articles about economic and financial concepts, NMF created many granular, specific topics about economic and financial matters. Economic and financial concepts existed in a minor subset of articles in the corpus, but they were sufficiently different from the content in other articles so that topics were formed from these distinct patterns. Similarly, NMF tended to form topics out of

specific events, which also can be thought of as distinct patterns.

It is unclear whether the actual algorithmic differences between NMF and LDA cause these differences, or if the tf-idf versus term frequency representation is the major cause. Judging from the nature of tf-idf (explained in section 2.3.1) and previous research (Chen et al. 2019), the different characteristics between the two methods are believed to partly originate from the tf-idf representation of NMF. For the algorithmic differences between the two, it is descriptive to think of the underlying mechanisms of the NMF and LDA algorithms. NMF is a dimensionality reduction technique, aimed to find a lower subspace that accurately describes the most significant and diverse patterns in the data, limited by the number of dimensions or topics it is allowed to factorize the corpus into. LDA is a generative method that samples common terms in a corpus, conditioned on terms that frequently occur together to create representative topics that mirrors the corpus in a probabilistic sense. This thesis does not attempt to pinpoint the exact cause of the differences between NMF and LDA, but rather highlight how their resulting topic models differ.

The number of topics to extract by a topic model is highly dependent on the use case. It is possible to find an optimal number of topics by optimizing quantitative metrics, but it is important to remember that these metrics aim to simulate human interpretability. A low number of topics to find will result in broader topics, while higher values will give more granular topics. The main finding in this thesis regarding the number of topics to find is that practitioners need to determine where along this granularity spectrum that the topic granularity should be. This aligns with the reasoning by Stevens et al. (2012) and Jacobi, Van Atteveldt, and Welbers (2016). In a quantitative sense, NMF peaks in performance at a low number of topics as can be seen in figure 6.3, while Gibbs sampling LDA peaks at around 40 topics as can be seen in Figure 6.4. The qualitative results only present topics from models created to find 40 topics but the number of topics to find is highly related to the use case. Practicioners should first decide what level of granularity of the topics that is desired and then, optionally, decide a suitable range of K for the desired granularity and employ quantitative measures to find an optimal value in that range. When observing the topic granularity for different values of K during the algorithm selection phase, a K value between 30 and 80 appeared to yield the most meaningful results as a categorization framework for Swedish news articles. This is a general recommendation from the results of this thesis for Swedish news articles and confirms similar findings by Stevens et al. (2012) and Jacobi, Van Atteveldt, and Welbers (2016) for English articles.

To summarize the answer to the above research question, data preparation is paramount when using a topic model. Dimensionality reduction techniques such as stopword removal, removal of infrequent terms, and lemmatization, especially for Swedish, should be utilized. Furthermore, filtering out only nouns in the corpus is judged to give the best results for a categorization framework. Proper nouns can lead to issues and misleading topics, however,

and there are strong arguments for using only common nouns in topic models for article categorization. The use of ngrams can give enhanced expressiveness in topic models but can lead to unwanted results due to a lack of control, and its benefits depend on the data representation that is used. NMF and LDA are both suitable topic modeling algorithms for categorizing news articles, but with different characteristics. NMF finds specific patterns in the data and can give more diverse, granular topics in general, while LDA creates a broader representation in its topics. The cause for these differences most likely lie in both the data representation and the algorithms. Therefore, there is no optimal algorithm as it all depends on the particular use case, and this coincides with previous research (Grimmer and Stewart 2013). The number of topics to find is highly dependent on the use case, but 30 to 80 topics is a general recommendation for categorizing Swedish news articles to yield coherent, meaningful results.

How well do topic model algorithms generalize to news article datasets of multiple news brands?

The evaluated topic model algorithms in this thesis were found to generalize well when learned with datasets made up of multiple news brands in the sense that topics did not degrade quantitatively, and that they were still meaningfully interpretable in a qualitative sense. Furthermore, NMF and LDA algorithms behave differently in this regard. These differences are perhaps more clearly illustrated when applying the algorithms to datasets of varying content from different brands.

Topic models are, like all machine learning methods, heavily data dependent, meaning that the characteristics of the data determine the characteristics of the topic model. It can be said that the intra-topic semantics is determined by the dataset. This means that concepts related within topics are determined by how these concepts co-occur in articles, and previous research presented in chapter 3 does indicate this. For example, Chandelier et al. (2018) used topic models to reveal perspectives on wolf recolonization in France through news articles, and thus highlighted public opinions about wolves, and what opinions co-occur in articles. Similarly, Q. Liu et al. (2019) explored concepts, such as illnesses, that are mentioned together with third-hand smoke in two corpora of Chinese and American news articles and highlight what illnesses that are co-related in these corpora. The results in this thesis also show this sort of relations, such as topics that connect *kvinna (woman)* to *våld (violence)*, which makes clear that these terms are statistically correlated to some degree. The term *kina (china)* was often related to *coronavirus (coronavirus)*. This is clearly an artifact due to the datasets' timespan up to February 2020, as the Covid19 pandemic originated in China at around that time, without having significantly spread to other countries. It can also be said that the inter-topic distribution is determined by the dataset. This means that the span of concepts that all topics from a topic model cover and their diversity is determined by how diverse the articles in the corpus are. For example, Chandelier et al. (2018) explored not only what

opinions are related within topics, but also what opinions existed overall in the corpus related to wolf recolonization in France. Q. Liu et al. (2019) not only explored how illnesses and third-hand smoke are related within topics but also overarchingly what illnesses are covered within the corpora. The results in this thesis highlight similar inter-topic distribution variations. The models learned from datasets with a high amount of economic-related articles in them tended to form a large number of granular topics on economic matters and thus reducing the conceptual span of the model. The models that were learned from a diverse set of articles generally tended to learn a broader span of concepts, with less granular topics on specific concepts.

The NMF and LDA algorithms both generalized well but yielded different outcomes due to characteristic differences between these algorithms as described earlier. NMF finds more specific patterns and forms many more economic related topics compared to LDA on a corpus biased towards economics. The usefulness of these economic-specific topics on other articles in the dataset unrelated to economics was however very limited. LDA formed broader topics that were better applicable across the entire corpus, but some of these topics tended to become incoherent. Again, both methods are useful as a categorization framework and as a content analysis tool for news articles, but their usefulness is determined by the use case.

How do the topics created by topic modeling differ from other categorization methods for Swedish news articles, and what are the strengths and weaknesses of topic modeling compared to these?

Given the results, it is clear that topic modeling is a powerful method for deriving insights from enormous sources of news article text data. Topic modeling, by its attribute of being an unsupervised machine learning method, avoids the major drawback of requiring predefined labels, as prominent supervised NLP models do. Supervised models such as BERT (Devlin et al. 2018) or ELMo (Peters et al. 2018) excel in tasks such as text classification, but they often require significantly more costly resources in the form of labeled data to learn a model compared to topic models (Grimmer and Stewart 2013; Quinn et al. 2010). Topic modeling was in this thesis feasible to apply on large volumes of data to learn a model relatively fast, mostly due to the reason for using unlabeled data as input. As discussed earlier, data preprocessing is instrumental for the success of a topic model, but this task was substantially time-consuming in relation to the actual modeling. Topic models use unlabeled data as input which is one of its key strengths, but this can also be a major weakness as it gives uncontrollable models. In chapter 3, Jacobi, Van Atteveldt, and Welbers (2016) emphasize the lack of control by showcasing that topics can allude to a variety of different concepts and that it is impossible to control if the topic model should find topics that relate to a specific concept, an event, or a geographic location. A similar lack of control was also shown by the experiments in this thesis. The topics, both for LDA and NMF, had different meanings and sometimes alluded to general concepts such as healthcare, whilst sometimes described a certain event or the

concept of a geographic location. It is clear that there is no certainty of semantic conformity in the taxonomic system created by topic models. In comparison, supervised methods and especially manual methods are more controllable.

A result of this thesis is that both automated methods as well as manual methods have clear strengths and weaknesses respectively. It is clear that topic modeling is at best an approximation of the vastly more complex nature of language as understood by humans and sometimes yields incorrect topics. This was not a surprising result since it aligns with previous research presented in chapter 3, for example, Grimmer and Stewart (2013). This is partly understood as a consequence of the representation problems in limitations of data with respect to language as previously discussed, partly due to the topic model algorithm. There are several examples in the topics shown in the topics presented in the results that have little semantic meaningfulness, and to say that an article has a theme that relates to such a topic would have no meaning at all.

The findings in the results confirm the conclusion drawn by Grimmer and Stewart (ibid.), that automated methods should be augmentations of manual analysis, but do not replace it. Topic modeling did not show the validity to create topics that are intended to be shown for an end consumer. The reason is the mentioned incoherent topics, but also the risk of the model creating unwanted topics, for example, topics being ethically questioned. There is no obvious example of this in the topics presented in this thesis, but such topics are a possibility. This ethical aspect is a compelling argument that topic modeling is not suited for categorizing articles and presenting these categorizations to an end consumer. However, it is not only topic models that can create topics that can be ethically questionable, as this can also be a problem for manual methods due to the risk of human bias and subjectivity. The result of this thesis leaves no answer to the question of what methods can be used to minimize these ethical risks and instead leaves this question open to further research and discussions within news media organizations.

As mentioned in the introduction, there exists a manual method in form of a manual tagging system at Bonnier News today. The brands categorize their news articles manually, with sections and tags. The different brands have different such sections and tags, hence there exists no manual uniform categorization framework. The journalists define what section an article belongs to, for example, “Ekonomi” (Economy), and add a few tags, for example, “BNP” (gross national product) and “Finansminister” (Minister of Finance). The sections the article belongs to is an umbrella term like “Sport” (Sport), “Kultur” (Arts), “Nöje” (Entertainment), “Klimat” (Climate), Ledare (Editorial), etc. These sections are stable and do not change often. Hence, these sections are a well-established way to categorize articles by tacit knowledge of which articles belong to what section. On the other hand, their robustness may cause problems for articles that span a variety of concepts. For example, articles about climate would naturally be defined as ”Klimat” (Climate), but often these articles relate to

other sections since the climate issue is something that is talked about in all parts of our society. The result is that these articles also could have been defined as, for example, "Nöje" (Entertainment) or "Ekonomi" (Economy). Topic modeling does also suffer from this problem, but the result in the qualitative evaluation indicated that topic modeling might handle these cases in a better way than the manual sections because of the multiple terms in topics.

The tags also have their strengths and weaknesses compared to topic modeling. They somewhat mitigate the problem mentioned above, since a reader can filter out articles from a variety of sections, such as the tag "Facebook", with articles from for example "Ekonomi" (Economy), "Kultur" (Arts) as well as Ledare (Editorial). These are naturally more granular than the sections since a journalist can choose several tags for an article, or create new tags if so desired. A consequence of this is that the number of tags can be inflated. For DN only, there are around 1400 different tags to the published articles over a 30 day period. Hence, tags are useful for a reader interested in a specific concept, but are less useful for a broad categorization framework of articles or serve as a basis to understand general reader behaviors. As mentioned earlier in this discussion, the hyperparameter K decides the level of granularity that the topic modeling should result in. A low number of topics to find will result in broader topics, while higher values will give more granular topics. Hence, topic modeling poses a huge advantage compared to the sections, but even more the tags, which is the ability to regulate the number of categories.

Another important aspect in the comparison between topic modeling and the existing manual tagging system is that topic models produce topics consisting of several terms, instead of one term. In the presented topics in the qualitative evaluation results section, five terms were found reasonable to present from a presentation point of view. These are the top five most probable terms for a topic of all terms in the corpus, but topic modeling enables choosing this number as fit. If this characteristic is a strength or weaknesses is understood as highly use case related.

A major strength of topic modeling compared to manual methods is the amount of data it can analyze. Manual methods performing some kind of content analysis might even be impossible due to enormous amounts of data since human-based methods are time and resource-intensive. Chandelier et al. (2018) argued that topic modeling possibly could be better to find hidden patterns and relationships between articles since the method generates topics independently from human preconceptions. This confirms the result shown in this thesis since many of the topics were quite unexpected but valuable results for understanding the content in the articles. This result also confirms their idea that topic modeling is a suitable tool for investigating trends and to analyze variations in media content such as coverage of a specific issue.

To summarize, the results showed that topics created by topic models are, compared to manual methods, more unreliable which is an effect of its uncontrollable nature. This is one of the major weaknesses of topic modeling. Its strength on the other hand is the capability

to analyze large amounts of text with a low startup cost which is the uniqueness of topic modeling. Moreover, the results confirmed previous studies arguing that topic modeling is a powerful method for deriving insights from enormous sources of news article text data but is not able to replace manual methods. This is especially the case if the result is intended to be shown for an end consumer. This suggests that topic modeling can be used as a complement to other methods for content analysis or categorization, confirming previous studies (Grimmer and Stewart 2013) presented in chapter 3.

Is topic modeling a viable framework for categorization of Swedish news articles from multiple brands and what are its main advantages and limitations?
The results indicate that topic modeling is a comprehensive method to use as a basis for a uniform categorization framework for Swedish news articles. The NMF and LDA algorithms are both deemed suitable, but the two methods have different characteristics when applied to multiple brands. Hence choosing which method most adequate to use as a basis for such a framework is use case dependent. However, topic modeling as implemented in this thesis was not found suitable to serve as a basis for a categorization framework intended to be shown for an end consumer. This means that topic modeling should not replace a manual tagging system, but that it instead can be appropriately used to create a categorization framework for internal use in an organization. The major strength of such a framework is the possibility to create a uniform taxonomy for articles within the whole organization with a low startup cost.

The advantages and limitations of topic models in general have been discussed earlier in this chapter. The advantages and limitations of a topic modeling categorization framework will be further explored in relation to the use case within Bonnier News, which is presented in chapter nine. Chapter 9 deals with the implementation of a topic model to create a uniform categorization framework within Bonnier News. The outlined implementation enables content analysis for all articles that are produced by every brand within the whole media group, which was difficult to do prior to this thesis.

8 Conclusion

Throughout the experiments in this thesis, multiple influencing choices have had to be made with regards to finding an optimal topic model for categorizing Swedish news articles. This thesis highlights in part that data selection and data preparation consist of at least half the work in topic modeling applications. It is difficult to evaluate topic models objectively due to multiple hyperparameters in topic modeling algorithms as well as in data preparation, and that the evaluation methods are tightly coupled with human judgment. The main conclusion is that it is difficult to continuously assess all variables as one hyperparameter is changed and that some choices have to be fixed in the evaluation process. An enabling success factor is thus a strong and flexible data preparation framework. Furthermore, the data on which topics models are learned is the main determinant of their usefulness. Input corpora with a high number of articles on a specific subject or with specific opinions in them will yield topics that resemble those subjects or opinions. Nouns are considered to yield the most meaningful topics for Swedish news articles, of which proper nouns greatly boost the potential for misleading topics due to misclassifications, leaving only common nouns as a more solid choice for a categorization framework.

The NMF and LDA algorithms have different characteristics in their outputs and create different topic distributions on the same datasets. NMF tends to find and represent specific patterns in the dataset to capture more variations, while LDA tends to find and represent more holistic topics to better mirror the entire dataset. The algorithms thus have different strengths and weaknesses depending on the use case and there is no optimal choice between the two. The same can be said for the number of topics to find, as the level of granularity in topics is most often more important than higher quantitative metric scores.

In comparison with other categorization techniques for Swedish news articles, topic modeling is unique in its capability to derive insights from large amounts of text with a low startup cost. However, it is clear that the resulting topics are more unreliable and uncontrollable than for example the result of a human-based method. The conclusion is that topic modeling is a powerful method for deriving insights from enormous sources of news article text data but is not able to replace manual methods. These findings confirm the conclusions by previous studies, that automated methods such as topic modeling should be augmentations of manual analysis and categorization, but do not replace it.

Topic modeling with both NMF and LDA can be comprehensive methods to use as a basis for a categorization framework for Swedish news articles from multiple brands. However, the topics are not suitable to be presented for end consumers of the news articles. Its major strength is instead the possibility to align different brands into one uniform categorization framework, that enables content analysis over news articles spanning multiple brands.

8.1 Future research

The conclusions in this thesis do leave some unanswered questions, and the most important of them are presented in this section.

The tf-idf representation is, in this thesis, thought to be a major factor in the differences between NMF and LDA. The degree to which the tf-idf representation determines these differences is unclear. It can therefore be interesting to apply NMF and LDA to similar data representations, to better understand which of these differences originate from the tf-idf representation, and what originates from the algorithms. Furthermore, the bag-of-words representation used by NMF and LDA are incomplete, and other data representations that better capture term sequences are interesting to evaluate for enhanced topic modeling. This can also include better data preparation methods, such as being able to form more usable ngrams.

The RS metric presented in this thesis offer a new evaluation property of topic models. The usefulness, implications and validity of using this metric is an unexplored area within topic modeling evaluation. This thesis also offers a qualitative topic analysis, where the characteristics of NMF and LDA are discussed by human interpretations of their resulting topics. This has been done in previous research but these works often focus on analyzing topics from a single model, while not contrasting different topic models and their topics against each other. Even though the human interpretation of topics as an evaluation method is highly subjective, further research in this direction can help to better understand the characteristics of topic modeling algorithms. Further research by similar methods as in this thesis, or by more evolved qualitative topic evaluation on a larger scale is thus strongly warranted. This includes examining the limitations of using a qualitative analysis approach as employed in this thesis, to assess the validity of this method. A related research endeavor is to interpret topics as one scales the number of topics to find, to qualitatively examine how topic granularity varies. A unique aspect of this thesis is the use of articles in Swedish. This could be an explanation of the contrasting findings to previous research, but this is not shown in this thesis. The application and analysis of topic modeling algorithms on Swedish text compared to English is therefore strongly warranted.

Lastly, the experiments in this thesis used NMF and LDA implementations by two software libraries with extensive but basic hyperparameter configurations. The use of other libraries with different configurations of topic models could potentially lead to more successful outcomes. The hyperparameter selection for these implementations was done manually. Automated tools for hyperparameter searches could potentially give better results in future evaluations of topic models. Furthermore, Chen et al. (2019) note that external knowledge can be used to enhance topic models if suitable external knowledge exists for the use case. A news media context has many external knowledge sources, such as article sections or tags, and the use of this knowledge could enhance topic models used in a news context. Future research on modified topic

models is therefore motivated, and on other methods apart from topic models that fulfill the same purpose, to benchmark similar-purpose methods against each other.

9 Implementation

This thesis was a collaboration with the organization Bonnier News (BN). The authors have mainly been in contact with the data analytics team at Dagens Nyheter (DN) and the machine learning team at BN, who assisted with great use case perspectives, domain knowledge of Swedish news media, and technical expertise. This chapter will present this collaboration in more detail, starting with how the use case was established, how the final algorithm implemented in production was chosen, as well as the process of implementing the code in the production environment. Finally, this chapter will present the value creation, both what has been done within the scope of this thesis, but also possibilities for further value creation.

9.1 Use Case

At the beginning of this thesis project, the usability of topic models within BN was unclear. Topic modeling had previously mostly been used with English news articles, hence topic models' capacity to create meaningful results with Swedish news articles was not acknowledged. To address this issue, the main purpose of this thesis was aimed to evaluate this. Moreover, at the beginning of this project, there was no confirmed use case for topic modeling at BN. To address this, a workshop was carried out. The purpose of the workshop was to discuss the possible use cases for topic modeling in a creative and explorative way. Participants in the workshop were domain experts within BN, such as data analysts, data scientists, product owners, etc. Among several interesting ideas discussed, one of the most valuable ideas was understood as ideas regarding cross-brand activities.

As mentioned in the previous chapter, the uniqueness of topic modeling enables cross-brand analysis. This idea builds on the ability to run topic modeling with data from all brands within BN. The result from such topic modeling, the topic distribution over documents and term distribution over topics, is intended to be used in two, separate but alike, use cases. The first use case is the creation of a machine learning feature, the second is content monitoring within the whole organization.

9.1.1 Machine Learning Feature

In their daily business, BN applies several machine learning models in order to provide information to enhance cross-brand sales, cross-brand marketing, and optimizing the distribution of single-copy newspapers. In some of these cases, the feature of what topic interests their readers is a meaningful feature to use to predict future behavior. Previously, this prediction has been using the sections from the different brands as an input variable. The prospect is that this is a feature that a topic model can provide, not bound to a specific brand and instead uniform across the organization.

9.1.2 Content Monitoring

The other use case, also related to cross-brand activities, is the content monitoring case. This means the use of content analysis that is performed on all of BN's content, instead of only analyzing content within each brand as is previously done. The derived insight from such a content analysis could for example enable possibilities to identify and avoid overlapping publications between the brands and instead recommendations for cross-brand publications. The content monitoring is suggested to be visualized in a dashboard, making it as easy and understandable as possible for data analysts to derive insights.

9.2 Model in Production

The topic modeling algorithm chosen to be used in production was NMF optimized for the Frobenius norm, with the multiplicative update (MU) solver, non-regularized. This decision was based on the result presented and discussed in this thesis. In summary, this algorithm had the advantages of being fast and achieving high coherence scores, especially for the chosen number of topics (explained further on in this section), and was more practical to implement within the IT infrastructure at BN. NMF also showed the advantage of being better capable of creating topics that mapped individual articles to a single topic, which was considered as a beneficial property. It should be noted that this was not an obvious choice and that LDA was also considered as a strong candidate to be used.

The only part of speech (POS) class chosen to be used for the production implementation was common nouns. Therefore, all tokens apart from common nouns were filtered out during the production preprocessing. Other preprocessing steps, such as lemmatization, removal of stopwords, etc. was employed in the same way as presented in chapter 5. The choice to use only common nouns was done due to the problems regarding proper nouns in topics models and especially so for NMF, discussed in chapter 7. An important aspect of this decision came from the case-specific goal to implement a topic model algorithm that could be used at a later stage to relearn a model that finds a similar topic distribution, hence being more stable in the topics that are discovered. Names of persons, events, etc. (often proper nouns) that are mentioned in news articles change over time, while concepts (often common nouns) remain. The reason for relearning a model at a later stage is that a topic model learned on news articles from a specific time period will create a topic distribution that mirrors the news coverage of that period. General news coverage is likely to be very similar from day to day, but given enough time the distribution of concepts covered may change. It could therefore be useful to be able to relearn a model after a longer period of time.

The NMF algorithm was implemented in BN's IT infrastructure and integrated with their data warehouse for articles, article metadata, and other data. The algorithm was applied on a dataset of 100 000 articles for the same brands as in the BN dataset presented in section 4.3 from the time period 2019-01-01 until 2020-03-31. The distribution of the brands was

slightly different from the BN dataset used in the experiments of this thesis, but as a whole, it was intended to capture the same distribution of content for local newspapers, industry magazines, etc. as was done in the BN dataset. The resulting topic model was then stored in the IT infrastructure to be used for classification purposes of unseen articles.

The number of topics was chosen as 30, a choice based on coherence score, suitability for the use case, and stability. For both use cases, a number of topics between 30 and 40 were found as a reasonable amount, partly to visualize in a dashboard, partly to divide reader interests. After settling that around 30 to 40 topics would fit the use cases, the chosen model was run over a range of K with step size 1 to find if there was a local optimum in coherence score close to these numbers. One optimum turned out to be at 30 topics. These topics are found in Appendix D.

The main purpose of the learned topic model was to be used as a classification tool with the ability to classify all articles within BN into the topic model specified topic categorization. The stored learned model was therefore reused for classifying newly published articles. The same data preprocessing was applied to new unseen articles to be classified as the data used by the algorithm for learning the model. A scheduling program was then implemented to periodically query newly published articles from the data warehouse, process and classify them by the topic model, and then insert the resulting topic metadata of those articles back into the data warehouse. This process was also manually done on older articles spanning back a certain period of time for analysis purposes of recent times. The metadata output of the topic model classification of an article consists of the article identifier, the topic name (i.e. the terms of one of the K topics found), and the topic score (i.e. the numeric score of that topic from the topic model). Each article received a metadata entry for each topic, to retain information about what topics that one article had been classified into and to what degree. By querying the data warehouse and aggregating on these scores, a user could find the top n topics for a given article and a measure of how well each of those n topics described that article. This was used for querying topics as machine learning features and as a data source for visualizing the topic results for articles from a given period, which was done in BN's data visualization environments.

9.3 Value Creation

The implementation of topic modeling within BN has successfully added value in both use cases mentioned in section 9.1, but also opened up possibilities for further value creation.

The created topics' ability to be used in the existing machine learning models at BN will be evaluated further by BN's machine learning team, to see if the topics should be used as an input variable instead of the sections used today. The problem with the sections is partly that they differ between the brands, partly that they are historical artifacts of "how news should be categorized", hence not mirroring the themes in Swedish news articles in the best

way nowadays. There is potential that the topics can be successful in both these regards. Although, there exist questions regarding how often the model should be updated to best mirror themes written in Swedish news. In addition to evaluating the topics as a feature in the existing machine learning models, the machine learning team at BN sees big potential in this kind of method. For example to be used in other ways in their models or future models, such as to describe a user's reader interest by a topic distribution.

BN conducts detailed content monitoring for their specific brands. For example, they have dashboards that visualize the number of published articles, page views, etc. in a given timespan by the tags, which enable one kind of content analysis and monitoring of trends in the reader's interests. However, this is within each brand. What the implementation of topic modeling has enabled is the possibility for them to "zoom-out" and perform content monitoring within the whole organization. For this purpose, dashboards were created for further insights into if BN writes about what they want/believe/should, and what topics interest BN's readers. Also, an overall increased understanding of what is written on different brands was visualized to enhance initiatives for cross-brand publications. In Figure 9.1, 9.2, and 9.3 a few examples from these dashboards are presented as screenshots.

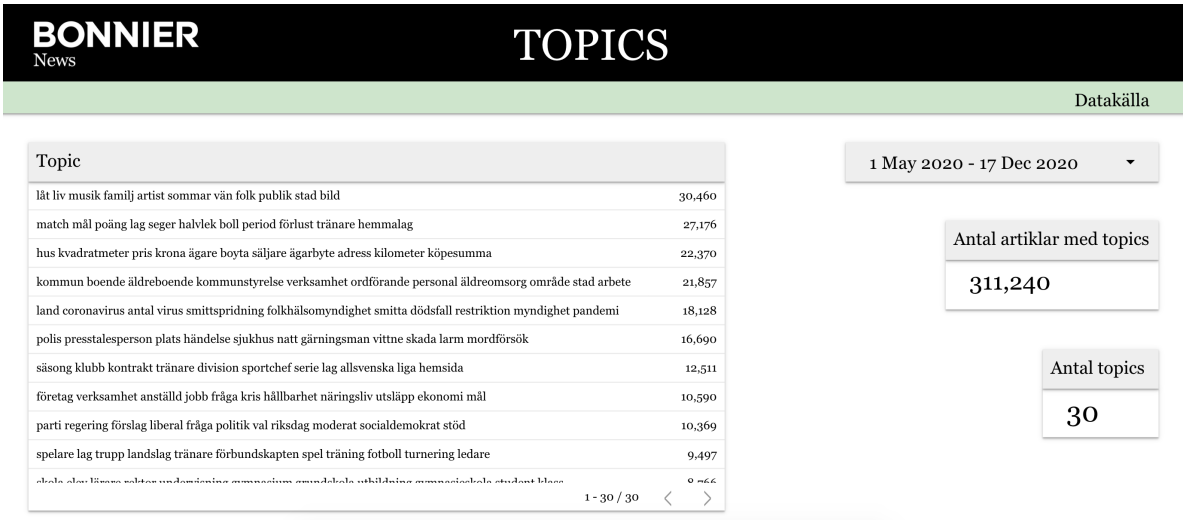

Figure 9.1: The data source page of the dashboard, presenting the topics, the number of articles within a date range that is categorized with a topic, and the total number of topics.

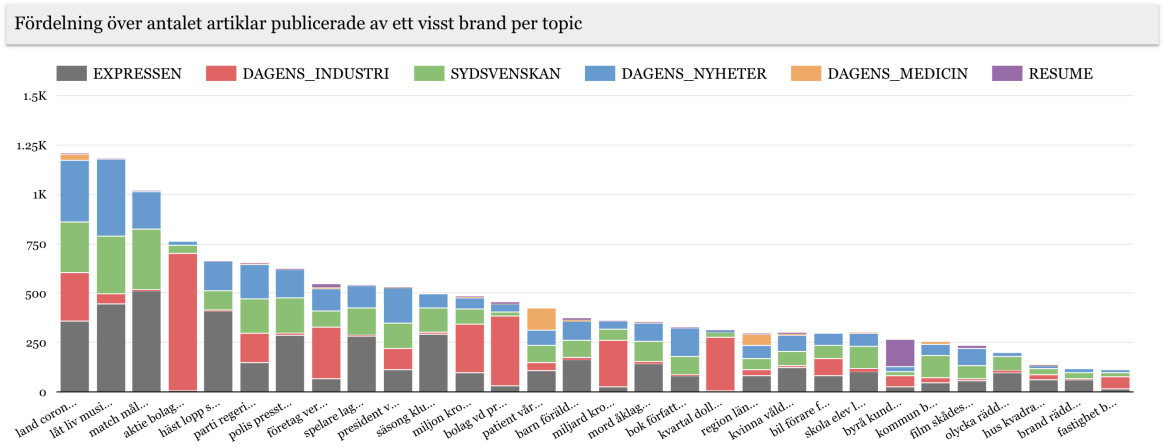

Figure 9.2: A bar chart presenting the distribution of articles published by six different brands per topic.

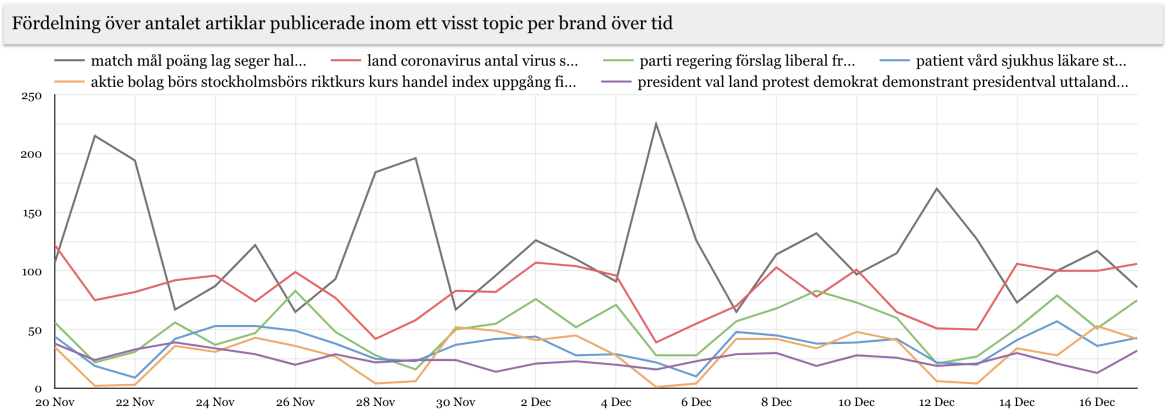

Figure 9.3: A time series over the distribution of articles published in sex different topics by all brands within Bonnier News.

Topic modeling enables other value-creating possibilities, not specific to the two use cases. As mentioned in previous research in section 3.2 and the discussion in chapter 7, topic modeling can be used to analyze variations in media content such as coverage of specific issues. As an example, this can be done by running a topic modeling algorithm on only climate articles, for example, filtered by the section "Klimat" (Climate) at DN, and evaluating what areas are covered by this issue. Another example is to run a topic modeling algorithm only on articles related to a political party (or parties) to see what concepts are covered with a certain party. Other value creation possibilities are to give the reader recommendation of articles within a topic of the user's interest or other kinds of personalizations based on the topics. The use cases for topic modeling are many, and its usage for various internal and analytical purposes by media organizations is clearly motivated.

References

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. Springer.

Blei, David M (2012). "Probabilistic topic models". In: *Communications of the ACM* 55.4, pp. 77–84.

Blei, David M and John D Lafferty (2009). "Topic models". In: *Text mining: classification, clustering, and applications* 10.71, p. 34.

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

Bonnier News (Aug. 2020). *Vi når och berör fler*. `https://www.bonniernews.se/`, Last accessed 2020-12-02.

Casalino, Gabriella, Nicoletta Del Buono, and Corrado Mencar (2016). "Nonnegative matrix factorizations for intelligent data analysis". In: *Non-negative Matrix Factorization Techniques*. Springer, pp. 49–74.

Caswell, David (2019). "Structured journalism and the semantic units of news". In: *Digital Journalism* 7.8, pp. 1134–1156.

Chandelier, Marie et al. (2018). "Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling". In: *Biological conservation* 220, pp. 254–261.

Chang, Jonathan et al. (2009). "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems* 22, pp. 288–296.

Chen, Yong et al. (2019). "Experimental explorations on short text topic mining between LDA and NMF based Schemes". In: *Knowledge-Based Systems* 163, pp. 1–13.

Chi, Eric C and Tamara G Kolda (2012). "On tensors, sparsity, and nonnegative factorizations". In: *SIAM Journal on Matrix Analysis and Applications* 33.4, pp. 1272–1299.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, pp. 391–407.

Devlin, Jacob et al. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Du, Rundong et al. (2017). "DC-NMF: nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modeling". In: *Journal of Global Optimization* 68.4, pp. 777–798.

Feldman, Ronen, James Sanger, et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Gillis, Nicolas (2014). "The why and how of nonnegative matrix factorization". In: *Regularization, optimization, kernels, and support vector machines* 12.257, pp. 257–291.

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1, pp. 5228–5235.

Grimmer, Justin and Gary King (2011). "General purpose computer-assisted clustering and conceptualization". In: *Proceedings of the National Academy of Sciences* 108.7, pp. 2643–2650.

Grimmer, Justin and Brandon M Stewart (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". In: *Political analysis* 21.3, pp. 267–297.

Haselmayer, Martin and Marcelo Jenny (2014). "Measuring the tonality of negative campaigning: Combining a dictionary approach with crowd-coding". In: *Political context matters: Content analysis in the social sciences. University of Mannheim.*

Hedlund, Turid, Ari Pirkola, and Kalervo Järvelin (2001). "Aspects of Swedish morphology and semantics from the perspective of mono-and cross-language information retrieval". In: *Information Processing & Management* 37.1, pp. 147–161.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010). "Online learning for latent dirichlet allocation". In: *advances in neural information processing systems*, pp. 856–864.

Hofmann, Thomas (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.

Hsieh, Cho-Jui and Inderjit S Dhillon (2011). "Fast coordinate descent methods with variable selection for non-negative matrix factorization". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1064–1072.

Jacobi, Carina, Wouter Van Atteveldt, and Kasper Welbers (2016). "Quantitative analysis of large amounts of journalistic texts using topic modelling". In: *Digital Journalism* 4.1, pp. 89–106.

Lau, Jey Han, David Newman, and Timothy Baldwin (2014). "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539.

Lee, Daniel D and H Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". In: *Nature* 401.6755, pp. 788–791.

Lindsey, Robert et al. (2007). "Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness". In: *8th International Conference of Cognitive Modeling, ICCM.*

Liu, Lin et al. (2016). "An overview of topic modeling and its current applications in bioinformatics". In: *SpringerPlus* 5.1, p. 1608.

Liu, Qian et al. (2019). "Data analysis and visualization of newspaper articles on thirdhand smoke: a topic modeling approach". In: *JMIR medical informatics* 7.1, e12414.

M'sik, Ben and Ben Msik Casablanca (2020). "Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus". In: *International Journal* 9.4.

Martin, Fiona and Mark Johnson (2015). "More efficient topic modelling through a noun only approach". In: *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 111–115.

McCallum, Andrew Kachites (2002). "Mallet: A machine learning for language toolkit". In: *http://mallet. cs. umass. edu*.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26, pp. 3111–3119.

Mimno, David et al. (2011). "Optimizing semantic coherence in topic models". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Neuendorf, Kimberly A and Anup Kumar (2015). "Content analysis". In: *The international encyclopedia of political communication*, pp. 1–10.

Östling, Robert (2018). "Part of Speech Tagging: Shallow or Deep Learning?" In: *Northern European Journal of Language Technology* 5, pp. 1–15.

– (2020). *efselab*. `https://https://github.com/robertostling/efselab`.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peters, Matthew E et al. (2018). "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365*.

Qaiser, Shahzad and Ramsha Ali (2018). "Text mining: use of TF-IDF to examine the relevance of words to documents". In: *International Journal of Computer Applications* 181.1, pp. 25–29.

Quinn, Kevin M et al. (2010). "How to analyze political attention with minimal assumptions and costs". In: *American Journal of Political Science* 54.1, pp. 209–228.

Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408.

Sarkar, Dipanjan (2019). *Text analytics with Python: a practitioner's guide to natural language processing*. Apress.

Seung, D and L Lee (2001). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems* 13, pp. 556–562.

Stevens, Keith et al. (2012). "Exploring topic coherence over many models and many topics". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961.

Steyvers, M and T Griffiths (2007). "Probabilistic topic models. Handbook of Latent Semantic Analysis. Edited by TK Landauer, DS McNamara, S. Dennis, W. Kintsch". In: *NJ: Erlbaum*.

Information Science in Korea using Topic Modeling." Journal of the Korean Society for Information Management 30.1, pp. 7–32.

Straka, Milan (2018). "UDPipe 2.0 prototype at CoNLL 2018 UD shared task". In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 197–207.

Suri, Pranav and Nihar Ranjan Roy (2017). "Comparison between LDA & NMF for event-detection from large text stream data". In: *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*. IEEE, pp. 1–5.

Surjandari, Isti et al. (2018). "Mining Web Log Data for News Topic Modeling Using Latent Dirichlet Allocation". In: *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, pp. 331–335.

Xu, Wei, Xin Liu, and Yihong Gong (2003). "Document clustering based on non-negative matrix factorization". In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 267–273.

Appendices

A Dataset Types

Dataset type: DN (single-brand)
Dagens Nyheter

Dataset type: DN/Di/HD (multi-brand)
One third of each: Dagens Nyheter, Dagens Industri, Helsingborgs Dagblad

Dataset type: BN (multi-brand)
One third, large morning papers: Dagens Nyheter, Sydsvenskan

One third, industry magazines: Aktuell Hållbarhet, Dagens Industri, Dagens Media, Dagens Medicin, Fastighetsnytt, Resumé

One third, local newspapers: Helsingborgs Dagblad, Falköpings Tidning, JönköpingsPosten, Jnytt, Skövde Nyheter, Skaraborgs Läns Tidning, Smålands Dagblad, SmålandsTidningen, Smålänningen, Tranås Tidning, VetlandaPosten, Värnamo Nyheter, VästgötaBladet

Dataset type: Extrinsic data
Dagens Nyheter, Sydsvenskan, Aktuell Hållbarhet, Dagens Industri, Dagens Media, Dagens Medicin, Fastighetsnytt, Resumé, Helsingborgs Dagblad, Falköpings Tidning, Jönköpings-Posten, Jnytt, Skaraborgs Läns Tidning, Skövde Nyheter, Smålands Dagblad, Smålands-Tidningen, Smålänningen, Tranås Tidning, VetlandaPosten, Värnamo Nyheter, Västgöta-Bladet

B POS Classes and Stopwords

Part of speech (POS) classes removed

Adjective (ADJ), verb (VERB), adverb (ADV), subordinating conjunction (SCONJ), auxiliary verb (AUX), punctuation (PUNCT), adposition (ADP), pronoun (PRON), determiner (DET), particle (PART), coordinating conjunction (CCONJ), numeral (NUM), interjection (INTJ).

Stopwords removed

Swedish stopwords = tjugoen, fyra, gärna, ha, ändå, långsam, komma, tar, går, både, gått, bara, oss, man, sjuttonde, ska, båda, rätt, ligger, gör, ur, alltid, men, tack, fått, dagen, alltså, bort, sextio, fall, upp, andra, eller, följande, godast, hellre, mest, behövde, överst, övre, som, mina, mer, eftersom, inför, med, hade, tolfte, förlåt, nittio, rakt, sätt, viktig, enligt, ta, beslutat, någonting, ger, likställda, annat, femtio, tjungo, redan, längre, knappast, vänstra, mittemot, åtta, dom, ingenting, nitton, allas, de, hela, noll, mycket, tror, snart, sedan, några, enkla, långsammast, för, första, nedersta, lättare, bli, nej, ofta, säger, likställd, ibland, varken, ut, vara, sagt, dit, heller, sina, nödvändig, tills, gälla, procent, dessa, beslutit, var, godare, åttionde, fanns, delen, inget, kvar, den, ner, dagar, långt, åttonde, imorgon, hit, är, själv, femton, hon, efter, nittionde, kunnat, också, verkligen, tro, alla, sitt, dina, elva, dig, enkel, sen, lilla, står, och, möjligen, blev, tog, ser, nödvändigt, igår, högst, er, vilka, sex, kunde, varför, allt, nödvändigtvis, inga, olikt, våra, fin, tjugotre, ännu, nionde, tredje, samma, nummer, helst, adertonde, trettionde, ni, hennes, varsågod, tre, gäller, fast, liten, litet, fem, tillsammans, slutligen, gjorde, idag, andras, mellan, bäst, ett, blivit, innan, fjärde, nödvändiga, sent, mig, in, sade, något, hur, hundraett, en, honom, senast, små, tidigt, vad, ditt, tjugoett, vårt, mitt, värre, dagarna, stort, än, över, borta, ja, fyrtio, sjunde, utanför, trettio, på, vem, bra, här, dess, sjutton, han, genast, inom, om, varit, nittonde, stor, sig, sista, hjälp, jämfört, trettonde, kom, gjort, vår, bättre, vidare, högre, min, god, varifrån, gick, nu, stora, måste, fram, möjlig, nog, gång, hög, vänster, sjätte, tvåhundra, bådas, borde, jag, sexton, siste, finnas, nr, större, sjuttionde, sist, genom, bakom, höger, kan, längst, tolv, då, göra, kommer, att, dag, artonde, enkelt, denna, kommit, gott, så, även, många, två, hundra, era, få, oftast, femtonde, goda, därför, blir, långsamt, tio, femte, viktigt, tjugotvå, mindre, ert, femtionde, nästa, säga, någon, nån, tjugonde, tidig, dock, möjligtvis, lägga, fick, lite, framför, henne, till, din, senare, annan, gällt, samt, fyrtionde, nåt, vi, nio, flesta, helt, ursäkt, vilken, dem, tionde, skall, vet, olika, heter, mera, långsammare, kanske, vems, åtminstone, fler, från, skulle, viktigare, behövas, ge, kr, möjligt, elfte, sämst, sextionde, ute, lätt, tjugo, beslut, nedre, flera, gå, lika, nya, väl, inuti, tidigare, inte, sextonde, sämre, haft, kunna, kolla, legat, adjö, aderton, nederst, ned, förra, före, inne, hans, tretton, sjuttio, ligga, minst, finns, mot, deras, ettusen, viktigast, smått, det, åttio, tidigast, där, varje, menar, har, visst, du, övermorgon, all, sju, behöva, bland, störst, i, ingen, vid, del, kör, sin, lättast, utan, aldrig,

vilket, behövt, artonn, får, av, när, igen, hundraen, vill, fjortonde, året, under, skriver, fjorton, detta, länge, vart

Above is standard words, below are words we added:
åt, vare, vars, vore, gav, apropå, sådant, ju, valt, se, sett, hör, ses, säg, knappt, drygt, låter, stund, medan, låg, sa, låg, ställs, stå, övrig, ab, gånger, 1980-tal, år, ej, f.n., kl., år, kl, åring, sätt, person, årsålder, dag, tid, gång, sak, minut, sekund, timme, nent, människa, del, de, fl, sd, mp, kd, quot, ;s, vecka, månad

English stopwords = i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, his, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, don't, should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, didn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ma, mightn, mightn't, mustn, mustn't, needn, needn't, shan, shan't, shouldn, shouldn't, wasn, wasn't, weren, weren't, won, won't, wouldn, wouldn't

C Topics

NMF Topics from DN Data

Topics	English translations
A1. barn, förälder, familj, mamma, sverige,	child, parent, family, mother, sweden
A2. polis, plats, presstalesperson, händelse, larm,	police, location, press spokesperson, incident, alarm
A3. trump, president, donald, usa, demokrat	trump, president, donald, usa, demokrat
A4. may, brexit, storbritannien, eu, theresa	may, brexit, britain, eu, theresa
A5. match, mål, lag, spelare, seger	match, goal, team, player, victory
A6. kina, coronavirus, virus, usa, wuhan	china, coronavirus, virus, usa, wuhan
A7. parti, löfven, liberal, väljare, politik	party, löfven, liberal, voter, politics
A8. kvinna, våldtäkt, våld, lägenhet, brott	woman, rape, violence, apartment, crime
A9. iran, usa, irak, soleimani, attack	iran, usa, iraq, soleimani, attack
A10. hongkong, demonstrant, protest, demonstration, carrie	hongkong, protester, protest, demonstration, carrie
A11. brand, räddningstjänst, australien, skogsbrand, byggnad	fire, rescue service, australia, forest fire, building
A12. skola, elev, lärare, skolinspektion, rektor	school, student, teacher, school inspectorate, principal
A13. kommun, regering, förslag, miljard, arbetsförmedling	municipality, government, proposal, billion, employment service
A14. krona, miljon, miljard, kvartal, bolag	krona, million, billion, quarter, company
A15. turkiet, syrien, erdogan, kurd, idlib	turkey, syria, erdogan, kurd, idlib
A16. thunberg, greta, klimataktivist, klimat, värld	thunberg, greta, climate activist, climate, world
A17. film, bok, liv, författare, roman	movie, book, life, author, novel
A18. bil, volvo, elbil, mil, förare	car, volvo, electric car, mile, driver
A19. johnson, boris, premiärminister, labour, nyval	johnson, boris, prime minister, labor, re-election
A20. patient, vård, läkare, region, sjukvård	patient, care, doctor, region, healthcare
A21. nordkorea, kim, usa, jong_un, sydkorea	north korea, kim, usa, jong_un, south korea
A22. mord, åklagare, tingsrätt, brott, fängelse	murder, prosecutor, district court, crime, prison
A23. land, president, attack, val, stad	country, president, attack, election, city
A24. öberg, vm, karlsson, frida, johaug	öberg, world championship, karlsson, frida, johaug
A25. bank, swedbank, penningtvätt, kund, granskning	bank, swedbank, money laundering, customer, review
A26. sjukhus, ambulans, mordförsök, skada, pojke	hospital, ambulance, attempted murder, injury, boy
A27. venezuela, maduro, guaidó, juan, nicolás	venezuela, maduro, guaidó, juan, nicolás
A28. biden, sander, joe, demokrat, bernie	biden, sander, joe, democrat, bernie
A29. tåg, trafik, olycka, stockholm, trafikverket	train, traffic, accident, stockholm, swedish transport administration
A30. aik, hammarby, klubb, säsong, malmö	aik, hammarby, club, season, malmö
A31. djurgården, frölunda, period, mål, match	djurgården, frölunda, period, goal, match
A32. ryssland, ukraina, putin, zelenskyj, president	russia, ukraine, putin, zelenskyj, president
A33. eu, europa, storbritannien, land, leyen	eu, europe, britain, country, leyen
A34. göteborg, stad, ifk, stockholm, kommun	gothenburg, city, ifk, stockholm, municipality
A35. israel, netanyahu, palestinier, västbanken, benjamin	israel, netanyahu, palestinians, the west bank, benjamin
A36. indien, pakistan, kashmir, modi, delhi	india, pakistan, kashmir, modi, delhi
A37. plan, boeing, pilot, flygplan, sas	plane, boeing, pilot, aircraft, sas
A38. sverige, asap, rocky, norge, artist	sweden, asap, rocky, norway, artist
A39. prins, meghan, harry, drottning, andrew	prince, meghan, harry, queen, andrew
A40. meter, sjöström, sarah, frisim, final	meters, sjöström, sarah, freestyle swimming, final

NMF Topics from DN/Di/HD Data

Topics	English translations
B1. aktie, bolag, krona, miljon, stockholmsbörs	share, company, krona, million, stockholm stock exchange
B2. trump, president, donald, usa, demokrat	trump, president, donald, usa, democrat
B3. polis, plats, presstalesperson, händelse, sjukhus	police, location, press spokesperson, incident, hospital
B4. hus, kvadratmeter, pris, krona, ägare	house, square meters, price, krona, owner
B5. parti, liberal, regering, val, sverige	party, liberal, government, election, sweden
B6. match, mål, lag, säsong, poäng	match, goal, team, season, point
B7. miljon, krona, kvartal, resultat, skatt	million, krona, quarter, profit, tax
B8. johnson, boris, brexit, storbritannien, eu	johnson, boris, brexit, britain, eu
B9. riktkurs, krona, rekommendation, behåll, köp	target price, krona, recommendation, keep, purchase
B10. bank, swedbank, penningtvätt, nordea, kund	bank, swedbank, money laundering, nordea, customer
B11. dollar, miljard, bloomberg, intäkt, aktie	dollars, billion, bloomberg, revenue, stock
B12. barn, förälder, familj, mamma, liv	children, parent, family, mother, life
B13. kina, coronavirus, virus, tull, handelskrig	china, coronavirus, virus, customs, trade war
B14. kvinna, våldtäkt, våld, brott, misshandel	woman, rape, violence, crime, assault
B15. inköpschefsindex, index, industri, tjänstesektor, markit	purchasing manager index, index, industry, service sector, markit
B16. bil, elbil, förare, väg, tesla	car, electric car, driver, road, tesla
B17. iran, usa, attack, sanktion, irak	iran, usa, attack, sanction, iraq
B18. kvartal, mkr, rapport, rörelseresultat, resultat	quarter, mkr, report, operating profit, profit
B19. brand, räddningstjänst, eld, larm, byggnad	fire, rescue service, fire, alarm, building
B20. volvo, cars, geely, lastbil, kina	volvo, cars, geely, truck, china
B21. hongkong, demonstrant, protest, demonstration, carrie	hong kong, protester, protest, demonstration, carrie
B22. månad, försäljning, industriproduktion, detaljhandel, statistik	month, sales, industrial production, retail, statistics
B23. skola, kommun, elev, lärare, stad	school, municipality, student, teacher, city
B24. ränta, riksbanken, centralbank, inflation, fed	interest rate, riksbanken, central bank, inflation, fed
B25. eu, storbritannien, land, sverige, eu_kommission	eu, britain, country, sweden, european commission
B26. miljard, krona, överskott, underskott, vinst	billion, krona, margin, loss, profit
B27. bolag, vd, företag, marknad, kund	company, ceo, company, market, customer
B28. olycka, sjukhus, ambulans, väg, lastbil	accident, hospital, ambulance, road, truck
B29. sverige, vm, norge, tävling, öberg	sweden, world championship, norway, competition, öberg
B30. turkiet, land, syrien, ryssland, president	turkey, country, syria, russia, president
B31. mord, åklagare, tingsrätt, brott, fängelse	murder, prosecutor, district court, crime, prison
B32. usa, nordkorea, kim, jong_un, sydkorea	usa, north korea, kim, jong_un, south korea
B33. euro, miljon, analytiker, rörelseresultat, prognos	euro, million, analyst, operating profit, forecast
B34. börs, index, wall, street, dow	stock exchange, index, wall, street, dow
B35. film, quot, regissör, låt, skådespelare	movie, quot, director, song, actor
B36. fat, oljepris, lager, vecka, olja	barrels, oil price, storage, week, oil
B37. may, theresa, parlament, brexit, premiärminister	may, theresa, parliament, brexit, prime minister
B38. thunberg, greta, klimat, värld, klimataktivist	thunberg, greta, climate, world, climate activist
B39. huawei, ericsson, usa, företag, nokia	huawei, ericsson, usa, company, nokia
B40. löfven, stefan, statsminister, regering, kristersson	löfven, stefan, prime minister, government, kristersson

NMF Topics from BN Data

Topics	English translations
C1. sverige, land, regering, fråga, värld	sweden, country, government, question, world
C2. polis, presstalesperson, plats, händelse, natt	police, press spokesperson, place, incident, night
C3. match, mål, lag, poäng, seger	match, goal, team, points, victory
C4. krona, miljon, kvartal, resultat, bolag	krona, million, quarter, profit, company
C5. kommun, kommunstyrelse, krona, verksamhet, förskola	municipality, municipal board, krona, function, preschool
C6. bil, olycka, räddningstjänst, väg, ambulans	car, accident, rescue service, road, ambulance
C7. patient, läkare, behandling, läkemedel, vård	patient, doctor, treatment, medicine, care
C8. kvinna, misshandel, våldtäkt, barn, åtal	woman, assault, rape, child, prosecution
C9. kina, hongkong, usa, land, coronavirus	china, hongkong, usa, country, coronavirus
C10. barn, förälder, förskola, familj, mamma	children, parent, preschool, family, mother
C11. trump, president, usa, donald, demokrat	trump, president, usa, donald, democrat
C12. brand, räddningstjänst, larm, byggnad, plats	fire, rescue service, alarm, building, location
C13. studie, forskare, utsläpp, rapport, resultat	study, researcher, emissions, report, results
C14. facebook, annonsör, annons, google, användare	facebook, advertiser, ad, google, user
C15. elev, skola, lärare, rektor, skolinspektion	student, school, teacher, principal, school inspectorate
C16. aktie, bolag, börs, stockholmsbörs, kurs	share, company, stock exchange, stockholm stock exchange, rate
C17. fastighet, kvadratmeter, bolag, vd, redaktion	property, square meters, company, ceo, editorial office
C18. eu, storbritannien, brexit, johnson, may	eu, britain, brexit, johnson, may
C19. tingsrätt, brott, åklagare, fängelse, mord	district court, crime, prosecutor, prison, murder
C20. parti, liberal, val, väljare, löfven	party, liberal, election, voter, löfven
C21. säsong, klubb, spelare, lag, division	season, club, player, team, division
C22. tidning, media, metro, bonnier, mittmedia	newspaper, media, metro, bonnier, mittmedia
C23. region, län, stockholm, jönköping, skåne	region, county, stockholm, jönköping, skåne
C24. miljard, dollar, kvartal, analytiker, euro	billion, dollar, quarter, analyst, euro
C25. företag, butik, produkt, jönköping, verksamhet	company, store, product, jönköping, business
C26. byrå, kund, kampanj, varumärke, resumé	agency, customer, campaign, brand, resumé
C27. skövde, skara, ifk, skaraborg, falköping	skövde, skara, ifk, skaraborg, falköping
C28. bok, författare, roman, liv, berättelse	book, author, novel, life, story
C29. bank, swedbank, penningtvätt, kund, ränta	bank, swedbank, money laundering, customer, interest
C30. vecka, miljö, förordning, lag, hållbarhetsområde	week, environment, regulation, law, sustainability area
C31. malmö, stad, mff, ff, lund	malmö, city, mff, ff, lund
C32. sjukhus, skaraborg, universitetssjukhus, karolinska, medicin	hospital, skaraborg, university hospital, karolinska, medicine
C33. vård, sjukvård, hälsa, utredning, omsorg	care, medical care, health, investigation, welfare
C34. vetlanda, speedway, bk, wirebrand, boro	vetlanda, speedway, bk, wirebrand, boro
C35. tävling, låt, svt, final, vm	competition, song, svt, final, world championship
C36. iran, usa, land, attack, sanktion	iran, usa, country, attack, sanction
C37. film, quot, regissör, skådespelare, kampanj	movie, quot, director, actor, campaign
C38. bostad, lägenhet, hus, område, projekt	housing, apartment, house, area, project
C39. greta, thunberg, klimat, värld, klimataktivist	greta, thunberg, climate, world, climate activist
C40. smhi, län, jönköping, temperatur, varning	smhi, county, jönköping, temperature, warning

LDA Topics from DN Data

Topics	English translations
D1. sverige, norge, finland, frankrike, grupp	sweden, norway, finland, france, group
D2. tåg, väg, brand, olycka, klocka	train, road, fire, accident, clock
D3. bild, museum, stad, utställning, tal	picture, museum, city, exhibition, speech
D4. svar, fråga, samhälle, insändare, sverige	answer, question, society, letter to the editor, sweden
D5. land, sverige, kvinna, syrien, attack	country, sweden, woman, syria, attack
D6. kina, land, hongkong, virus, coronavirus	china, country, hongkong, virus, coronavirus
D7. match, mål, lag, period, säsong	match, goal, team, period, season
D8. plan, meter, os, flygplats, idrott	plane, meter, olympic games, airport, sports
D9. djur, skog, brand, område, mark	animal, forest, fire, area, land
D10. medium, tidning, bild, dn, journalist	medium, newspaper, picture, dn, journalist
D11. myndighet, dn, arbete, chef, jobb	authority, dn, work, manager, job
D12. stockholm, stad, göteborg, kommun, plats	stockholm, city, gothenburg, municipality, location
D13. lag, indien, mål, klubb, match	team, india, goal, club, match
D14. london, prins, larsson, vecka, drottning	london, prince, larsson, week, queen
D15. land, protest, president, demonstrant, ledare	country, protest, president, protester, leader
D16. kyrka, häst, kläder, namn, väg	church, horse, clothes, name, road
D17. värld, tal, greta, historia, thunberg	world, speech, greta, history, thunberg
D18. peng, krona, bank, ekonomi, kommun	money, krona, bank, economy, municipality
D19. vm, karlsson, tävling, lopp, dam	world championship, karlsson, competition, race, lady
D20. sverige, land, utsläpp, klimat, projekt	sweden, country, emissions, climate, project
D21. kvinna, brott, åklagare, fängelse, domstol	woman, crime, prosecutor, prison, court
D22. anna, peter, vecka, lars, johansson	anna, peter, week, lars, johansson
D23. miljon, krona, företag, bolag, miljard	million, krona, company, corporation, billion
D24. film, serie, roll, regi, skådespelare	movie, series, role, direction, actor
D25. svt, anders, johan, erik, program	svt, anders, johan, erik, program
D26. mat, hand, kött, vin, vatten	food, hand, meat, wine, water
D27. val, parti, röst, kandidat, väljare	election, party, vote, candidate, voter
D28. barn, skola, elev, förälder, lärare	children, school, student, parent, teacher
D29. musik, låt, artist, scen, band	music, song, artist, stage, band
D30. iran, usa, ryssland, turkiet, israel	iran, usa, russia, turkey, israel
D31. familj, liv, barn, vän, mamma	family, life, children, friend, mother
D32. polis, plats, händelse, mord, område	police, place, event, murder, area
D33. eu, storbritannien, johnson, brexit, boris	eu, britain, johnson, brexit, boris
D34. kvinna, universitet, studie, forskare, forskning	woman, university, study, researcher, research
D35. bil, krona, pris, mil, elbil	car, krona, price, mile, electric car
D36. sjukhus, vård, region, patient, läkare	hospital, care, region, patient, doctor
D37. match, mål, spelare, lag, aik	match, goal, player, team, aik
D38. trump, usa, president, donald, demokrat	trump, usa, president, donald, democrat
D39. bok, författare, roman, liv, berättelse	book, author, novel, life, story
D40. parti, regering, politik, stefan, liberal	party, government, politics, stefan, liberal

LDA Topics from DN/Di/HD Data

Topics	English translations
E1. musik, låt, artist, band, scen	music, song, artist, band, stage
E2. storbritannien, avtal, eu, brexit, johnson	britain, agreement, eu, brexit, johnson
E3. mat, restaurang, djur, häst, vin	food, restaurant, animal, horse, wine
E4. henrik, spel, par, larsson, slag	henrik, game, par, larsson, stroke
E5. vård, patient, studie, sjukhus, läkare	care, patient, study, hospital, doctor
E6. månad, vecka, januari, antal, februari	month, week, january, number, february
E7. match, mål, lag, säsong, period	match, goal, team, season, period
E8. karlsson, lopp, nilsson, tävling, säsong	karlsson, race, nilsson, competition, season
E9. iran, usa, land, ryssland, attack	iran, usa, country, russia, attack
E10. miljon, krona, kvartal, miljard, bolag	million, krona, quarter, billion, company
E11. familj, liv, vän, pappa, mamma	family, life, friend, father, mother
E12. kommun, stockholm, stad, göteborg, malmö	municipality, stockholm, city, gothenburg, malmö
E13. match, mål, klubb, lag, spelare	match, goal, club, team, player
E14. plan, flygplats, coronavirus, vecka, australien	plane, airport, coronavirus, week, australia
E15. hus, krona, helsingborg, pris, kvadratmeter	house, krona, helsingborg, price, square meters
E16. barn, skola, förälder, elev, lärare	children, school, parent, student, teacher
E17. bok, författare, historia, liv, roman	book, author, story, life, novel
E18. bolag, bank, vd, swedbank, styrelse	corporation, bank, CEO, swedbank, board
E19. medium, tidning, svt, bild, facebook	medium, newspaper, svt, picture, facebook
E20. ekonomi, ränta, tillväxt, riksbanken, prognos	economy, interest rates, growth, riksbanken, forecast
E21. sverige, fråga, samhälle, svar, problem	sweden, question, society, answer, problem
E22. företag, jobb, anställd, chef, verksamhet	company, job, employee, manager, business
E23. film, quot, roll, serie, skådespelare	movie, quot, role, series, actor
E24. kina, usa, dollar, miljard, tull	china, usa, dollar, billion, customs
E25. trump, usa, president, donald, demokrat	trump, usa, president, donald, democrat
E26. myndighet, uppgift, utredning, fråga, information	authority, task, investigation, question, information
E27. polis, plats, brand, händelse, skada	police, location, fire, incident, injury
E28. sverige, anders, johan, norge, danmark	sweden, anders, johan, norway, denmark
E29. krona, peng, bostad, butik, kund	krona, money, housing, store, customer
E30. tal, utställning, museum, bild, verk	speech, exhibition, museum, picture, work
E31. sverige, vm, lag, meter, os	sweden, world championship, team, meter, olympic games
E32. land, protest, hongkong, president, demonstrant	country, protest, hong kong, president, protester
E33. aktie, krona, bolag, riktkurs, börs	share, krona, company, target price, stock exchange
E34. eu, land, europa, tyskland, frankrike	eu, country, europe, germany, france
E35. bil, volvo, krona, fordon, väg	car, volvo, krona, vehicle, road
E36. vatten, tåg, väg, skog, område	water, train, road, forest, area
E37. kvinna, brott, åklagare, fängelse, domstol	woman, crime, prosecutor, prison, court
E38. sverige, utsläpp, värld, klimat, mål	sweden, emissions, world, climate, goal
E39. parti, val, röst, väljare, kandidat	party, election, vote, voter, candidate
E40. regering, parti, förslag, stefan, liberal	government, party, proposal, stefan, liberal

LDA Topics from BN Data

Topics	English translations
F1. krona, miljon, peng, miljard, resultat	krona, million, money, billion, result
F2. familj, liv, mamma, vän, pappa	family, life, mom, friend, dad
F3. barn, skola, elev, förälder, ungdom	children, school, student, parent, youth
F4. vatten, skog, plan, område, grad	water, forest, level, area, degree
F5. match, lag, säsong, mål, spelare	match, team, season, goal, player
F6. företag, kund, produkt, butik, marknad	company, customer, product, store, market
F7. myndighet, utredning, information, fråga, regel	authority, investigation, information, question, rule
F8. tidning, media, bonnier, expressen, journalist	newspaper, media, bonnier, expressen, journalist
F9. byrå, kampanj, kommunikation, kund, varumärke	agency, campaign, communication, customer, brand
F10. bild, tal, kyrka, utställning, museum	picture, speech, church, exhibition, museum
F11. bolag, fastighet, vd, pressmeddelande, bostad	company, property, ceo, press release, housing
F12. jobb, företag, arbete, chef, anställd	job, company, work, manager, employee
F13. vecka, månad, sommar, januari, slut	week, month, summer, january, end
F14. skövde, skara, förening, falköping, skaraborg	skövde, skara, association, falköping, skaraborg
F15. region, vård, patient, sjukhus, läkare	region, care, patient, hospital, doctor
F16. match, mål, spelare, lag, sverige	match, goal, player, team, sweden
F17. bil, väg, plats, brand, olycka	car, road, location, fire, accident
F18. kvinna, land, grupp, våld, attack	woman, country, group, violence, attack
F19. usa, trump, kina, president, land	usa, trump, china, president, country
F20. tävling, lopp, meter, final, vm	competition, race, meter, final, world championship
F21. mål, match, period, lag, säsong	goal, match, period, team, season
F22. bil, energi, teknik, volvo, el	car, energy, technology, volvo, el
F23. jönköping, län, eksjö, nässjö, värnamo	jönköping, county, eksjö, nässjö, värnamo
F24. facebook, medium, google, innehåll, tv4	facebook, medium, google, content, tv4
F25. film, musik, quot, låt, scen	movie, music, quot, song, scene
F26. stad, hus, område, plats, lokal	city, house, area, location, local
F27. parti, regering, stefan, fråga, politik	party, government, stefan, question, politics
F28. anna, johansson, fredrik, henrik, erik	anna, johansson, fredrik, henrik, erik
F29. eu, storbritannien, val, regering, land	eu, britain, election, government, country
F30. bok, författare, liv, berättelse, roman	book, author, life, story, novel
F31. mat, restaurang, jul, kött, vin	food, restaurant, christmas, meat, wine
F32. aktie, bolag, bank, dollar, kvartal	stock, company, bank, dollar, quarter
F33. polis, brott, kvinna, händelse, tingsrätt	police, crime, woman, incident, district court
F34. mål, utsläpp, företag, rapport, hållbarhet	goals, emissions, company, report, sustainability
F35. malmö, johan, andersson, anders, nilsson	malmö, johan, andersson, anders, nilsson
F36. fråga, problem, samhälle, exempel, svar	question, problem, society, example, answer
F37. stockholm, göteborg, peter, pris, lars	stockholm, göteborg, peter, price, lars
F38. sverige, land, antal, värld, svensk	sweden, country, number, world, swedish
F39. studie, forskare, risk, sjukdom, patient	study, researcher, risk, disease, patient
F40. kommun, förslag, verksamhet, ordförande, budget	municipality, proposal, activity, chairman, budget

D Topics in Production

Topics	English translations
låt, liv, musik, familj, artist	song, life, music, family, artist
match, mål, poäng, lag, seger	match, goal, points, team, victory
hus, kvadratmeter, pris, krona, ägare	house, square meters, price, krona, owner
kommun, boende, äldreboende, kommunstyrelse, verksamhet	municipality, housing, retirement home, municipal board, function
land, coronavirus, antal, virus, smittspridning	country, coronavirus, number, virus, spread of infection
polis, presstalesperson, plats, händelse, sjukhus	police, press spokesperson, location, event, hospital
säsong, klubb, kontrakt, tränare, division	season, club, contract, coach, division
företag, verksamhet, anställd, jobb, fråga	company, business, employee, job, question
parti, regering, förslag, liberal, fråga	party, government, proposal, liberal, question
spelare, lag, trupp, landslag, tränare	player, team, squad, national team, coach
skola, elev, lärare, rektor, undervisning	school, student, teacher, principal, teaching
olycka, räddningstjänst, ambulans, sjukhus, väg	accident, rescue service, ambulance, hospital, road
brand, räddningstjänst, larm, byggnad, plats	accident, rescue service, ambulance, hospital, road
barn, förälder, familj, mamma, förskola	children, parent, family, mother, preschool
mord, åklagare, brott, tingsrätt, fängelse	murder, prosecutor, crime, district court, prison
miljon, krona, omsättning, resultat, rörelseresultat	million, krona, sales, profit, operating profit
häst, lopp, seger, tävling, start	horse, race, victory, competition, start
bil, förare, fordon, elbil, körning	car, driver, vehicle, electric car, driving
region, län, hälsa, sjukvård, folkhälsomyndighet	region, county, health, healthcare, public health authority
kvinna, våldtäkt, våld, lägenhet, bostad	woman, rape, violence, apartment, housing
president, val, land, protest, demokrat	president, election, country, protest, democrat
patient, vård, sjukhus, läkare, studie	patient, care, hospital, doctor, study
bok, författare, roman, liv, berättelse	book, author, novel, life, story
bolag, vd, pressmeddelande, styrelse, media	company, ceo, press release, board, media
aktie, bolag, börs, stockholmsbörs, riktkurs	share, company, stock exchange, stockholm stock exchange, target price
byrå, kund, varumärke, kampanj, kommunikation	agency, customer, brand, campaign, communication
miljard, krona, dollar, bank, euro	billion, krona, dollar, bank, euro
fastighet, bostad, kvadratmeter, lägenhet, hyresgäst	property, housing, square meters, apartment, tenant
kvartal, dollar, rapport, resultat, mkr	quarter, dollars, report, profit, SEK million
film, skådespelare, regissör, premiär, serie	movie, actor, director, premiere, series